

人工智能代理

勘测多模态交互的前景

Zane Durante^{1†*}、Qiuyuan Huang^{2‡*}、Naoki Wake^{2*}、

Ran Gong^{3†}、Jae Sung Park^{4†}、Bidipta Sarkar^{1†}、Rohan Taori^{1†}、Yusuke Noda⁵、Demetri Terzopoulos³、Yejin Choi⁴、Katsushi Ikeuchi²、Hoi Vo⁵、Li Fei-Fei¹、Jianfeng Gao²

¹斯坦福大学；²微软研究院，雷德蒙德；

³加州大学洛杉矶分校；⁴华盛顿大学；⁵微软游戏

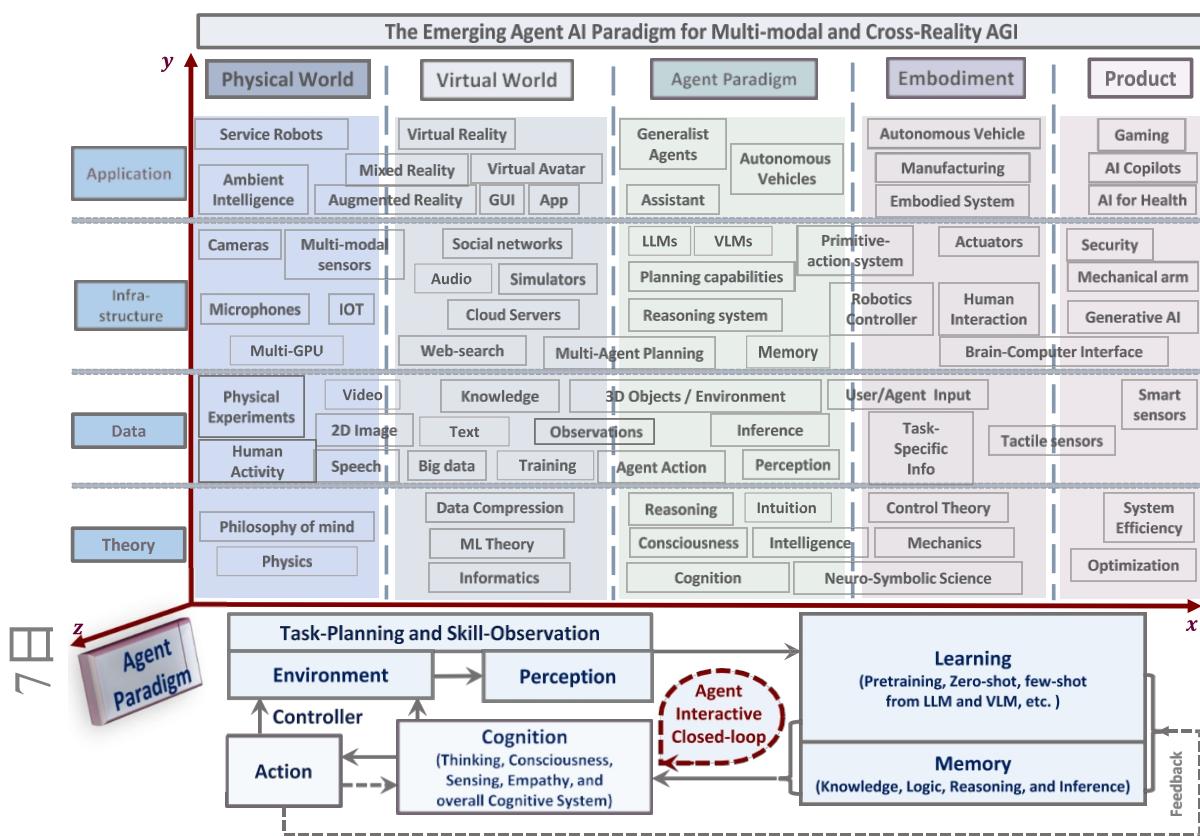


图 1：可在不同领域和应用中感知和行动的代理人工智能系统概览。Agent AI is emerging as a promising avenue toward Artificial General Intelligence (AGI). Agent AI training has demonstrated the capacity for multi-modal understanding in the physical world. It通过利用生成式人工智能和多个独立数据源，提供了一个与现实无关的训练框架。针对代理和行动相关任务训练的大型基础模型，在跨现实数据上进行训练后，可应用于物理和虚拟世界。我们介绍了一个代理人工智能系统的总体概况，该系统可以在许多不同的领域和应用中感知和行动，有可能成为使用代理范式实现 AGI 的途径。

*同等贡献。 ‡项目负责人。 †在雷德蒙德微软研究院实习期间完成的工作。

摘要

多模式人工智能系统很可能成为我们日常生活中无处不在的存在。让这些系统更具互动性的一个可行方法是在物理和虚拟环境中将它们体现为代理。目前，系统利用现有的基础模型作为创建具身代理的基本构件。将代理嵌入到这些环境中有助于提高模型处理和解释视觉及上下文数据的能力，这对于创建更复杂、更能感知上下文的人工智能系统至关重要。例如，一个能够感知用户行为、人类行为、环境物体、音频表达和场景的集体情绪的系统，可用于告知和指导代理在给定环境中的反应。为了加快对基于代理的多模态智能的研究，我们将“代理人工智能”定义为一类交互系统，它可以感知视觉刺激、语言输入和其他环境基础数据，并能与无限代理一起产生有意义的体现行动。特别是，我们探索的系统旨在通过结合外部知识、多感官输入和人类反馈，在下一个体现行动预测的基础上改进代理。我们认为，通过在基础环境中开发代理人工智能系统，还可以减轻大型基础模型的幻觉及其产生环境错误输出的倾向。新兴的代理人工智能领域涵盖了多模态交互中更广泛的体现和代理方面。除了在物理世界中行动和互动的代理之外，我们还设想在未来，人们可以轻松创建任何虚拟现实或模拟场景，并与虚拟环境中的代理进行互动。

目录

1 导言	5
1.1 激励	5
1.2 背景介绍	5
1.3 概述	6
2 代理人工智能集成	7
2.1 无限人工智能代理	7
2.2 使用大型基础模型的代理人工智能	8
2.2.1 幻觉	8
2.2.2 偏见与包容性	9
2.2.3 数据隐私与使用	10
2.2.4 可解释性和可说明性	11
2.2.5 推理增强	12
2.2.6 条例	13
2.3 面向新兴能力的代理人工智能	14
3 代理人工智能范式	15

勘测多模态交互的前景

3.1 法律硕士和文凭硕士	15
3.2 代理变压器定义	15
3.3 代理变形金刚制作	16
4 代理人工智能学习	17
4.1 战略和机制	17
4.1.1 强化学习 (RL)	17
4.1.2 模仿学习 (IL)	18
4.1.3 传统 RGB	18
4.1.4 情境学习	18
4.1.5 代理系统中的优化	18
4.2 代理系统（零发和少发级）	19
4.2.1 代理模块	19
4.2.2 代理基础设施	19
4.3 代理基础模型（预培训和微调级）	19
5 代理人工智能分类	20
5.1 通用代理领域	20
5.2 化身特工	20
5.2.1 行动代理	20
5.2.2 互动代理	21
5.3 模拟与环境代理	21
5.4 生成代理	21
5.4.1 AR/VR/Mixed-Rality 代理商	22
5.5 知识与逻辑推理代理	22
5.5.1 知识代理	23
5.5.2 逻辑代理	23
5.5.3 情感推理代理	23
5.5.4 神经符号制剂	24
5.6 法律硕士和文职法律硕士代理人	24
6 代理人工智能应用任务	24
6.1 游戏代理	24
6.1.1 NPC 行为	24
6.1.2 人与人之间的互动	25
6.1.3 基于代理的游戏分析	25
6.1.4 游戏场景合成	27
6.1.5 实验和结果	27

勘测多模态交互的前景

6.2 机器人.....	28
6.2.1 机器人学 LLM/VLM 代理。	30
6.2.2 实验和结果。	31
6.3 医疗保健.....	35
6.3.1 当前的医疗保健能力.....	36
6.4 多式联运代理.....	36
6.4.1 图像语言理解与生成.....	36
6.4.2 视频与语言理解和生成.....	37
6.4.3 实验和结果.....	39
6.5 视频语言实验.....	41
6.6 NLP 代理.....	45
6.6.1 法学硕士代理.....	45
6.6.2 普通法律硕士代理.....	45
6.6.3 遵循指示的法律硕士代理人	46
6.6.4 实验和结果.....	46
7 跨越模式、领域和现实的代理人工智能	48
7.1 促进跨模态理解的代理.....	48
7.2 跨域理解代理.....	48
7.3 跨模态和跨现实的交互式代理.....	49
7.4 模拟到真实的传输.....	49
8 代理人工智能的持续和自我完善	49
8.1 人机交互数据.....	49
8.2 地基模型生成的数据.....	50
9 代理数据集和排行榜	50
9.1 用于多代理游戏的 "CuisineWorld "数据集.....	50
9.1.1 基准.....	51
9.1.2 任务	51
9.1.3 衡量标准和评判.....	51
9.1.4 评估.....	51
9.2 音频-视频-语言预培训数据集。	51
10 更广泛的影响说明	52
11 伦理方面的考虑	53
12 多样性声明	53

参考文献	55
附录	69
A GPT-4V 代理提示详细信息	69
B 用于出血边缘的 GPT-4V	69
C 用于微软格斗模拟器的 GPT-4V	69
D 刺客信条奥德赛》的 GPT-4V	69
E 用于《战争机器 4》的 GPT-4V	69
F 星域 GPT-4V	75
作者传记	77
鸣谢	80

1 导言

1.1 动机

从历史上看，人工智能系统在 1956 年达特茅斯会议上被定义为能够从环境中收集信息并以有用的方式与之互动的人工生命形式。在这一定义的推动下，明斯基的麻省理工学院小组于 1970 年建立了一个名为 "复制演示" 的机器人系统，该系统可以观察 "积木世界" 的场景，并成功重建观察到的多面体积木结构。该系统由观察、规划和操纵模块组成，它揭示出这些子问题中的每一个都极具挑战性，因此有必要开展进一步的研究。人工智能领域分裂成若干专业子领域，这些子领域在解决这些问题和其他问题方面基本上各自取得了巨大进展，但过度还原论模糊了人工智能研究的总体目标。

要想超越现状，就必须回归以亚里士多德整体论为动机的人工智能基本原理。幸运的是，最近在大型语言模型（LLM）和可视语言模型（VLM）方面的革命使我们有可能创造出符合整体理想的新型人工智能代理。抓住这一机遇，本文探讨了将语言能力、视觉认知、上下文记忆、直觉推理和适应性融为一体的新模型。文章探讨了使用 LLM 和 VLM 完成这种整体综合的可能性。在探索过程中，我们还重新审视了基于亚里士多德 "最终原因" 的系统设计，即在前几轮人工智能发展中可能被忽视的目的论 "系统为何存在"。

随着功能强大的预训练 LLM 和 VLM 的出现，催化了自然语言处理和计算机视觉领域的复兴。现在，LLMs 在解读真实世界语言数据的细微差别方面表现出了令人印象深刻的能力，其能力往往与人类的专业知识并驾齐驱，甚至超越人类的专业知识（OpenAI, 2023 年）。最近，研究人员表明，LLMs 可以扩展为各种环境中的代理，在与特定领域的知识和模块配对后执行复杂的行动和任务（Xi 等人, 2023 年）。这些场景的特点是复杂的推理、对代理角色及其环境的理解，以及多步骤规划，考验代理在环境限制下做出高度细微和复杂决策的能力（Wu 等人, 2023 年；Meta Fundamental AI Research (FAIR) Diplomacy Team 等人, 2022 年）。

在这些初步努力的基础上，人工智能界正处于范式重大转变的风口浪尖，从创建用于被动、结构化任务的人工智能模型，过渡到能够在多样化的复杂环境中扮演动态代理角色的模型。在此背景下，本文研究了将 LLMs 和 VLMs 用作代理的巨大潜力，强调模型应融合语言能力、视觉认知、上下文记忆、直觉推理和适应能力。利用 LLMs 和 VLMs 作为代理，特别是在游戏、机器人和医疗保健等领域，不仅有望为最先进的人工智能系统提供一个严格的评估平台，还预示着以代理为中心的人工智能将对整个社会和行业产生变革性影响。如果得到充分利用，代理模型可以重新定义人类体验并提升操作标准。这些模式所带来的全面自动化潜力预示着各行各业和社会经济动态的巨大转变。正如我们将在第 11 节中阐述的那样，这种进步将与多方面的领先优势交织在一起，不仅是技术方面的，还有道德方面的。我们将深入探讨代理人工智能这些子领域的重叠部分，并在图 1 中说明它们之间的相互联系。

1.2 背景介绍

接下来，我们将介绍相关的研究论文，以支持代理人工智能的概念、理论背景和现代实现。

大型基础模型：LLMs 和 VLMs 一直在推动通用智能机器的发展（Bubeck 等人，2023 年；Mirchandani 等人，2023 年）。虽然它们是使用大型文本语料库进行训练的，但其卓越的问题解决能力并不局限于典型的语言处理领域。从数学推理（Imani et al., 2023; Wei et al., 2022; Zhu et al., 2022）到回答专业法律问题（Blair-Stanek et al., 2023; Choi et al., 2023; Nay, 2022），LLMs 都有可能解决以前认为只有人类专家或特定领域算法才能完成的复杂任务。最近的研究表明，可以使用 LLM 为机器人和游戏人工智能生成复杂的计划（Liang 等人，2022；Wang 等人，2023a,b；Yao 等人，2023a；Huang 等人，2023a），这是 LLM 作为通用智能代理的一个重要里程碑。

嵌入式人工智能：一些研究 利用 LLMs 来执行任务规划（Huang 等人，2022a；Wang 等人，2023b；Yao 等人，2023a；Li 等人，2023a），特别是 LLMs 的 WWW 规模领域知识和新兴的零拍嵌入式能力来执行复杂的任务规划和推理。最近的机器人研究也利用 LLMs 来执行任务规划（Ahn 等人，2022a；Huang 等人，2022b；Liang 等人，2022），方法是将自然语言指令分解为一系列子任务（自然语言形式或 Python 代码），然后使用底层控制器来执行这些子任务。此外，他们还结合环境反馈来提高任务性能（Huang 等人，2022b）、（Liang 等人，2022）和（Wang 等人，2023a）。

互动学习：为交互式学习而设计的人工智能代理采用机器学习技术和用户交互相结合的方式进行操作。最初，人工智能代理会在一个大型数据集上接受训练。这个数据集包括各种类型的信息，具体取决于代理的预期功能。例如，针对语言任务设计的人工智能将在大量文本数据的语料库上进行训练。训练涉及使用机器学习算法，其中可能包括神经网络等深度学习模型。这些训练模型能让人工智能识别模式、做出预测，并根据训练数据生成响应。人工智能代理还可以从与用户的实时互动中学习。这种互动学习有多种方式：1) 基于反馈的学习：人工智能根据用户的直接反馈调整其反应（Li 等人，2023b；Yu 等人，2023a；Parakh 等人，2023；Zha 等人，2023；Wake 等人，2023a,b,c）。例如，如果用户纠正了人工智能的反应，人工智能就可以利用这一信息改进未来的反应（Zha 等人，2023 年；Liu 等人，2023a）。2) 观察学习：人工智能观察用户的互动，并进行隐性学习。例如，如果用户经常问类似的问题或以特定的方式与人工智能互动，人工智能可能会调整自己的回应，以更好地适应这些模式。它允许人工智能代理理解和处理人类语言、多模型设置、解释跨现实语境并生成人类用户的回应。随着时间的推移，随着更多的用户互动和反馈，人工智能代理的性能一般会不断提高。这一过程通常由人类操作员或开发人员进行监督，确保人工智能学习得当，不会产生偏差或错误模式。

1.3 概述

多模态代理人工智能（MAA）是基于对多模态感官输入的理解，在给定环境中生成有效行动的一系列系统。随着大型语言模型（LLMs）和视觉语言模型（VLMs）的出现，从基础研究到应用领域已经提出了许多 MAA 系统。虽然这些研究领域通过与各领域的传统技术（如视觉问题解答和视觉语言导航）相结合而发展迅速，但它们在数据收集、基准测试和伦理视角等方面有着共同的兴趣。在本文中，我们将重点关注 MAA 的一些代表性研究领域，即多模态、游戏（VR/AR/MR）、机器人和医疗保健，并旨在就这些领域讨论的共同关注点提供全面的知识。因此，我们希望大家能够学习到 MAA 的基础知识，并获得进一步推动研究的真知灼见。具体学习成果包括

- MAA 概述：深入探讨其原理和在当代应用中的作用，让研究人员全面掌握其重要性和用途。
- 方法论：通过游戏、机器人技术和医疗保健领域的案例研究，详细举例说明 LLM 和 VLM 如何增强 MAA。
- 绩效评估：使用相关数据集对千年生态系统评估进行评估的指南，重点关注其有效性和普遍性。

勘测多模态交互的前景

- 伦理考虑：讨论部署人工智能代理的社会影响和道德标准，强调负责任的开发实践。
- 新兴趋势和未来排行榜：对各领域的最新发展进行分类，并讨论未来的发展方向。

基于计算机的行动和通用代理（GA）对许多任务都很有用。要想让通用代理对其用户真正有价值，它就必须能够自然地与各种环境和模式进行交互和泛化。我们的目标是培养一个充满活力的研究生态系统，并在代理人工智能社区中建立共同的认同感和目标感。MAA 有可能广泛应用于各种情境和模式，包括来自人类的输入。Therefore, we believe this Agent AI area can engage a diverse range of researchers, fostering a dynamic Agent AI community and

共同目标。在来自学术界和工业界的著名专家的带领下，我们希望本论文将是一次互动和丰富的体验，其中包括代理指导、案例研究、任务会议和实验讨论，确保为所有研究人员提供全面和引人入胜的学习体验。

本文旨在提供有关代理人工智能领域当前研究的一般和全面知识。为此，本文接下来的内容安排如下。第 2 节概述了人工智能如何从与相关新兴技术（尤其是大型基础模型）的整合中获益。第 3 节介绍了我们为训练 Agent AI 提出的新范式和框架。第 4 节概述了广泛用于训练人工智能的方法。第 5 节对各种类型的代理进行分类和讨论。第 6 节介绍了代理人工智能在游戏、机器人和医疗保健领域的应用。第 7 节探讨了研究界为开发多用途人工智能（Agent AI）所做的努力，这种人工智能能够应用于各种模式和领域，并弥合模拟与现实之间的差距。第 8 节讨论了代理人工智能的潜力，它不仅依赖于预先训练好的基础模型，还能通过与环境和用户的互动不断学习和自我完善。第 9 节介绍了我们为训练多模态人工智能（Agent AI）而设计的新数据集。第 11 节讨论了人工智能代理的伦理考量、局限性以及本文的社会影响等热点话题。

2 代理人工智能集成

以往研究中提出的基于 LLMs 和 VLMs 的基础模型在具身人工智能领域仍表现出有限的性能，特别是在理解、生成、编辑和在未见环境或场景中交互方面 (Huang 等人, 2023a; Zeng 等人, 2023)。因此，这些局限性会导致人工智能代理产生次优输出。当前以代理为中心的人工智能建模方法侧重于可直接访问且定义明确的数据（如世界状态的文本或字符串表示），通常使用从大规模预训练中学到的与领域和环境无关的模式来预测每个环境的行动输出 (Xi 等, 2023; Wang 等, 2023c; Gong 等, 2023a; Wu 等, 2023)。在 (Huang 等, 2023a) 一文中，我们结合大型基础模型，研究了知识引导下的协作和交互场景生成任务，结果表明基于知识的 LLM 代理可以提高二维和三维场景理解、生成和编辑以及其他人机交互的性能 (Huang 等, 2023a)。通过整合代理人工智能框架，大型基础模型能够更深入地理解用户输入，从而形成复杂的自适应人机交互系统。LLM 和 VLM 的新兴能力在生成式人工智能、体现式人工智能、多模型学习的知识增强、混合现实生成、文本到视觉编辑、游戏或机器人任务中 2D/3D 模拟的人机交互等方面发挥着无形的作用。代理人工智能 最近在基础模型方面取得的进展，为开启代理的通用智能提供了迫在眉睫的催化剂。大型行动模型或代理视觉语言模型为通用代理系统（如复杂环境中的规划、问题解决和学习）开辟了新的可能性。代理人工智能（Agent AI）是元宇宙（metaverse）的进一步测试，也是 AGI 的早期版本。

2.1 无限人工智能代理

人工智能代理有能力根据其训练和输入数据进行解释、预测和响应。虽然这些能力是先进的，并在不断改进，但重要的是要认识到它们的局限性以及它们所训练的基础数据的影响。人工智能代理系统一般具备以下能力：1) 预测建模：人工智能代理可以根据历史数据和趋势预测可能的结果或建议下一步措施。例如，它们可以预测文本的续写、问题的答案、机器人的下一步行动或场景的解决方案。2) 决策：在某些应用中，人工智能代理可以根据自己的推断做出决策。一般来说，人工智能代理会根据什么最有可能实现特定目标来做出决策。在推荐系统等人工智能应用中，人工智能代理可以根据对用户偏好的推断，决定推荐哪些产品或内容

勘测多模态交互的前景

。3) 处理模糊性：人工智能代理通常可以根据上下文和训练推断出最可能的解释，从而处理模棱两可的输入。不过，它们的能力受到训练数据和算法范围的限制。

4) 持续改进：虽然一些人工智能代理有能力从新数据和交互中学习，但许多大型语言模型在训练后并不会持续更新其知识库或内部表征。它们的推断通常仅基于上次训练更新前的可用数据。

我们在图 2 中展示了增强型交互式代理，用于多模态和跨现实的整合，并具有一种出现机制。人工智能代理需要为每项新任务收集大量的训练数据，这对于许多领域来说可能成本高昂或根本不可能。在本研究中，我们开发了一种无限代理，它可以学习从以下设备中传输记忆信息



图 2：用于跨现实 2D/3D 具身生成和编辑交互的多模型代理人工智能。

将一般基础模型（如 GPT-X、DALL-E）应用于新的领域或场景，以便在物理或虚拟世界中进行场景理解、生成和交互式编辑。

这种无限代理在机器人技术中的一个应用是 RoboGen (Wang 等人, 2023d)。在这项研究中，作者提出了一个管道，可以自主运行任务命题、环境生成和技能学习的循环。RoboGen 致力于将大型模型中蕴含的知识转移到机器人技术中。

2.2 具有大型基础模型的代理人工智能

最近的研究表明，大型基础模型在创建数据方面发挥着至关重要的作用，这些数据可作为在环境施加的限制条件下确定代理行动的基准。例如，将基础模型用于机器人操纵 (Black 等人, 2023 年; Ko 等人, 2023 年) 和导航 (Shah 等人, 2023a; Zhou 等人, 2023a)。举例来说，Black 等人采用图像编辑模型作为高级规划器，生成未来子目标的图像，从而指导低级策略 (Black 等人, 2023 年)。在机器人导航方面，Shah 等人提出了一个系统，利用 LLM 从文本中识别地标，利用 VLM 将这些地标与视觉输入相关联，通过自然语言指令增强导航能力 (Shah 等人, 2023a)。

人们对根据语言和环境因素生成有条件的人体动作也越来越感兴趣。一些人工智能系统已被提出来根据特定的语言指令生成动作和行动 (Kim 等人, 2023 年; Zhang 等人, 2022 年; Tevet 等人, 2022 年)，并适应各种 3D 场景 (Wang 等人, 2022a)。这些研究强调了生成模型在提高人工智能代理在不同场景下的适应性

和响应能力方面日益增强的能力。

2.2.1 幻觉

生成文本的代理通常容易产生幻觉，即生成的文本不合逻辑或不忠于所提供的源内容（Raunak 等人，2021 年；Maynez 等人，2020 年）。幻觉可分为内在和外在两类（Ji 等人，2023 年）。内在性幻觉是指与以下内容相矛盾的幻觉

而外在幻觉则是指生成的文本包含了源材料中原本没有的额外信息。

减少语言生成中的幻觉率的一些可行方法包括使用检索增强生成（Lewis等人，2020年；Shuster等人，2021年）或其他通过外部知识检索为自然语言输出提供基础的方法（Dziri等人，2021年；Peng等人，2023年）。一般来说，这些方法通过检索额外的源材料和提供机制来检查生成的反应与源材料之间的矛盾，从而增强语言生成。

在多模态代理系统中，VLMs 也会产生幻觉（Zhou 等人，2023b）。基于视觉的语言生成产生幻觉的一个常见原因是过度依赖训练数据中物体和视觉线索的共现（Rohrbach 等人，2018 年）。完全依赖预训练 LLM 或 VLM 并使用有限的特定环境微调的人工智能代理尤其容易产生幻觉，因为它们依赖预训练模型的内部知识库来生成动作，可能无法准确理解所处世界的动态。

2.2.2 偏见与包容性

基于 LLM 或 LMM（大型多模态模型）的人工智能代理，由于其设计和训练过程中固有的几个因素而存在偏差。在设计这些人工智能代理时，我们必须注意包容性，了解所有最终用户和利益相关者的需求。就人工智能代理而言，包容性指的是所采用的措施和原则，以确保代理的响应和互动对来自不同背景的广大用户具有包容性、尊重性和敏感性。我们在下文中列出了代理偏见和包容性的主要方面。

- **训练数据：**基础模型是通过从互联网上收集的大量文本数据（包括书籍、文章、网站和其他文本资源）进行训练的。这些数据通常反映了人类社会中存在的偏见，模型会在无意中学习并复制这些偏见。这包括与种族、性别、民族、宗教和其他个人属性相关的刻板印象、偏见和倾斜观点。特别是，通过对互联网数据（通常只有英文文本）的训练，模型会潜移默化地学习西方、受过教育、工业化、富裕和民主（WEIRD）社会的文化规范（Henrich et al.然而，我们必须认识到，人类创建的数据集不可能完全没有偏见，因为它们经常反映了社会偏见以及最初生成和/或编译数据的个人的倾向性。）。
- **历史和文化偏见：**人工智能模型是在来自不同内容的大型数据集上进行训练的。因此，训练数据通常包括来自不同文化的历史文本或资料。特别是，来自历史资料的训练数据可能包含代表特定社会文化规范、态度和偏见的冒犯性或贬损性语言。这可能会导致模型延续过时的刻板印象，或无法完全理解当代文化的转变和细微差别。
- **语言和语境限制：**语言模型可能难以理解和准确表达语言中的细微差别，如讽刺、幽默或文化指涉。这可能会导致某些语境下的误解或有偏差的反应。此外，纯文本数据无法捕捉口语的许多方面，这可能导致人类对语言的理解与模型对语言的理解脱节。
- **政策和指导方针：**人工智能代理根据严格的政策和指导方针运行，以确保公平性和包容性。例如，在生成图像时，有规则要求对人物进行多样化的描述，避免与种族、性别和其他属性相关的刻板印

象。

- **过度泛化：**这些模型倾向于根据训练数据中的模式生成反应。这可能会导致过度概括，即模型产生的反应似乎对某些群体抱有成见或作出宽泛的假设。
- **持续监控和更新：**对人工智能系统进行持续监控和更新，以解决任何新出现的偏见或包容性问题。在这一过程中，用户的反馈和人工智能伦理方面的持续研究发挥着至关重要的作用。

- **主流观点的放大：**由于训练数据通常包含更多来自主流文化或群体的内容，因此模型可能会更偏向于这些观点，从而有可能低估或误导少数群体的观点。
- **道德和包容性设计：**在设计人工智能工具时，应将道德考量和包容性作为核心原则。这包括尊重文化差异、促进多样性以及确保人工智能不会延续有害的刻板印象。
- **用户指南：**还指导用户如何以促进包容性和尊重的方式与人工智能互动。这包括避免提出可能导致有偏见或不适当输出的请求。此外，这还有助于减少模型从用户互动中学习有害物质。

尽管采取了这些措施，人工智能代理仍会表现出偏见。目前，人工智能代理研发工作的重点是进一步减少这些偏见，增强人工智能代理系统的包容性和公平性。减少偏见的努力：

- **多样化和包容性的培训数据：**努力在培训数据中纳入更加多样和包容的来源。
- **偏差检测和纠正：**正在进行的研究侧重于检测和纠正模型响应中的偏差。
- **道德准则和政策：**模式通常受道德准则和政策的约束，旨在减少偏见并确保尊重和包容的互动。
- **多元化代表：**确保人工智能代理生成的内容或提供的回应能够代表人类的各种经历、文化、种族和身份。这一点在图像生成或叙事构建等场景中尤为重要。
- **减少偏差：**积极努力减少人工智能响应中的偏见。这包括与种族、性别、年龄、残疾、性取向和其他个人特征有关的偏见。目标是提供公平、平衡的回复，使陈规定型观念或偏见不再延续。
- **文化敏感性：**人工智能的设计具有文化敏感性，承认并尊重文化规范、习俗和价值观的多样性。这包括理解和适当应对文化参照和细微差别。
- **可访问性：**确保不同能力的用户（包括残疾用户）都能使用人工智能代理。这可能涉及到加入一些功能，让有视觉、听觉、运动或认知障碍的人更容易进行交互。
- **基于语言的包容性：**提供对多种语言和方言的支持，以满足全球用户群的需求，并对语言内部的细微差别和差异保持敏感（Liu 等人，2023b）。
- **道德和尊重的互动：**在程序设计中，Agent 会以道德和尊重的方式与所有用户互动，避免做出可能被视为冒犯、有害或不尊重的回应。
- **用户反馈和调整：**结合用户反馈，不断提高人工智能代理的包容性和有效性。这包括从互动中学习，以便更好地了解和服务于不同的用户群。
- **遵守包容性准则：**遵守人工智能代理包容性的既定准则和标准，这些准则和标准通常由行业团体、道德委员会或监管机构制定。

尽管做出了这些努力，但重要的是要意识到回答中可能存在的偏见，并用批判性思维来解释它们。人工智能代理技术和道德实践的不断改进旨在逐渐减少这些偏见。人工智能代理包容性的总体目标之一是创建一个尊重所有用户的代理，无论其背景或身份如何，都可以使用。

2.2.3 数据隐私和使用

人工智能代理的一个关键伦理考虑因素涉及了解这些系统如何处理、存储和可能检索用户数据。我们将在下文讨论主要方面：

数据收集、使用和目的。在使用用户数据提高模型性能时，模型开发人员会访问人工智能代理在生产过程中与用户交互时收集的数据。有些系统允许用户通过用户账户或向服务提供商提出请求来查看自己的数据。了解人工智能代理在这些交互过程中收集了哪些数据非常重要。这可能包括文本输入、用户使用模式、个人偏好，有时还包括更敏感的个人信息。用户还应了解如何使用从交互中收集到的数据。如果由于某种原因，人工智能掌握了某个特定个人或群体的错误信息，那么一旦发现，用户应该有一个机制来帮助纠正。这对准确性和尊重所有用户和群体都很重要。检索和分析用户数据的常见用途包括改进用户交互、个性化响应和系统优化。对于开发人员来说，确保数据不被用于未经用户同意的目的（如未经请求的营销）是极其重要的。

存储和安全。开发人员应了解用户交互数据的存储位置，以及采取了哪些安全措施来防止未经授权的访问或泄露。这包括加密、安全服务器和数据保护协议。确定是否与第三方共享代理数据以及在何种条件下共享代理数据极为重要。这应该是透明的，通常需要用户同意。

数据删除和保留。同样重要的是，用户要了解用户数据会被存储多久，以及用户如何要求删除数据。许多数据保护法赋予用户被遗忘的权利，这意味着他们可以要求删除自己的数据。人工智能代理必须遵守数据保护法，如欧盟的 GDPR 或加州的 CCPA。这些法律规定了数据处理做法和用户对其个人数据的权利。

数据移植性和隐私政策。此外，开发人员必须制定人工智能代理的隐私政策，记录并向用户解释如何处理他们的数据。其中应详细说明数据收集、使用、存储和用户权利。开发人员应确保在收集数据（尤其是敏感信息）时征得用户同意。用户通常可以选择退出或限制他们提供的数据。在某些司法管辖区，用户甚至有权要求以可转移到其他服务提供商的格式提供一份数据副本。

匿名化。对于用于更广泛分析或人工智能训练的数据，最好进行匿名化处理，以保护个人身份。开发人员必须了解他们的人工智能代理如何在交互过程中检索和使用用户历史数据。这可能是为了实现个性化或提高响应相关性。

总之，要了解人工智能代理的数据隐私，就必须了解如何收集、使用、存储和保护用户数据，并确保用户了解自己在访问、更正和删除数据方面的权利。用户和人工智能代理对数据检索机制的认识，对于全面了解数据隐私也至关重要。

2.2.4 可解释性和可说明性

模仿学习 → 解耦。通常使用强化学习（RL）或模仿学习（IL）中的连续反馈回路来训练代理，从随机初始化的策略开始。然而，这种方法在陌生环境中获取初始奖励时会遇到 "排行榜" 问题，特别是当奖励稀少或只有在长步骤交互结束时才能获得奖励时。因此，更好的解决方案是使用通过 IL 训练的无限记忆代理，它可以从专家数据中学习策略，改善对未知环境空间与新兴基础设施的探索和利用，如图 3 所示。通过专家特征帮助代理更好地探索和利用未知环境空间。人工智能代理可以直接从专家数据中学习政策和新的范式流

程。

传统的 IL 是让代理模仿专家示范者的行为来学习策略。然而，直接学习专家策略并不总是最佳方法，因为代理可能无法很好地概括未见过的情况。为了解决这个问题，我们建议学习一个带有上下文提示或隐含奖励函数的代理，该函数可以捕捉专家行为的关键方面，如图 3 所示。这就为无限记忆代理提供了执行任务的物理世界行为数据，这些数据是从专家演示中学来的。这有助于克服现有的模仿学习弊端，如需要大量的专家数据，以及在复杂任务中可能出现的错误。代理人工智能背后的关键理念由两部分组成：1) 无限代理，收集物理世界专家示范的状态-行动对；2) 虚拟环境，模仿代理生成器。模仿代理生成模仿专家行为的行动，而代理则通过减少专家行动与所学策略生成的行动之间差异的损失函数，学习从状态到行动的策略映射。

解耦 → 泛化。代理不依赖于特定任务的奖励函数，而是从专家示范中学习，专家示范提供了一系列涵盖不同任务方面的状态-行动对。然后，代理通过模仿专家的行为来学习将状态映射到行动的策略。模仿学习中的“去耦合”指的是将学习过程与特定任务的奖励函数分离开来，从而使政策能够在不同任务中通用，而无需明确依赖特定任务的奖励函数。通过解耦，代理可以从专家的示范中学习，并学会一种能适应各种情况的策略。解耦使迁移学习成为可能，在这一过程中，在一个领域学习到的策略只需最小限度的微调就能适应其他领域。通过学习不与特定奖励函数挂钩的通用策略，代理可以利用在一项任务中获得的知识，在其他相关任务中表现出色。由于代理不依赖于特定的奖励函数，它可以适应奖励函数或环境的变化，而无需进行大量的再训练。这使得学习到的策略更加稳健，并可在不同环境中通用。这里所说的“解耦”是指在学习过程中将两个任务分开：学习奖励函数和学习最优策略。

泛化 → 新生行为。泛化（Generalization）解释了如何从较简单的组件或规则中产生新兴属性或行为。其关键在于确定支配系统行为的基本要素或规则，如单个神经元或基本算法。然后，通过观察这些简单的组成部分或规则是如何相互作用的。这些组成部分之间的相互作用会导致复杂行为的出现，而这些复杂行为是无法通过单独研究单个组成部分来预测的。在不同复杂程度上的泛化可以让系统学习到适用于这些复杂程度的一般原则，从而产生新的特性。这样，系统就能适应新的情况，从更简单的规则中产生更复杂的行为。此外，跨越不同复杂度水平的概括能力有利于知识从一个领域转移到另一个领域，这有助于系统在新的环境中适应复杂行为的出现。

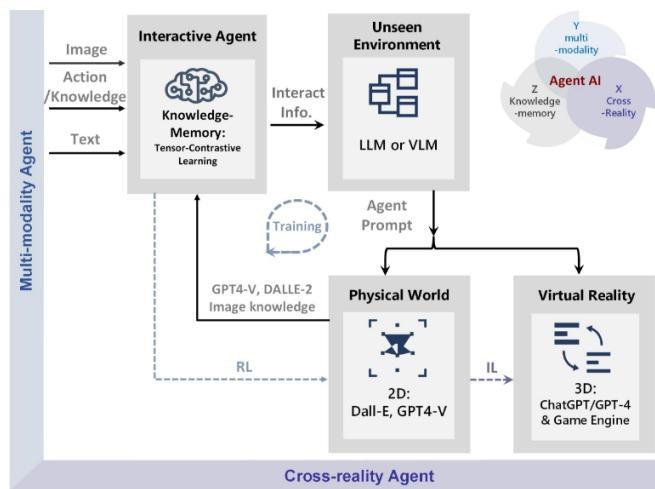


图 3：使用代理从候选人中识别与图像相关的文本的新互动机制示例。这项任务涉及使用来自网络的多模态人工智能代理和人类标注的知识交互样本来整合外部世界的信息。

2.2.5 推理增强

人工智能代理的推理能力在于其根据训练和输入数据进行解释、预测和响应的能力。虽然这些能力是先进的，并在不断改进，但重要的是要认识到它们的局限性以及它们所训练的基础数据的影响。特别是在大型语言模型中，它指的是根据所训练的数据和接收到的输入得出结论、进行预测和做出反应的能力。人工智能代理中的推理增强是指利用额外的工具、技术或数据来增强人工智能的自然推理能力，从而提高其性能、准确性和实用性。这在复杂的决策场景或处理细微或专业的内容时尤为重要。我们将在下文中指出推理增强的重要来源：

丰富数据。纳入额外的（通常是外部的）数据源以提供更多的上下文或背景，可以帮助人工智能代理做出更明智的推断，尤其是在其训练数据可能有限的领域。例如，人工智能代理可以从对话或文本的上下文中推断意义。它们会分析给定的信息，并利用这些信息来理解用户查询的意图和相关细节。这些模型精通识别数据中的模式。它们利用这种能力，根据在训练中学到的模式对语言、用户行为或其他相关现象进行推断。

算法增强。改进人工智能的底层算法，以做出更好的推断。这可能涉及使用更先进的机器学习模型，整合不同类型的人工智能（如将 NLP 与图像识别相结合），或更新算法以更好地处理复杂任务。语言模型推理涉及理解和生成人类语言。这包括把握细微差别，如语气、意图和不同语言结构的微妙之处。

人在回路（HITL）。在人工智能的推理过程中，人类的参与会起到至关重要的作用，例如在道德考量、创造性任务或模棱两可的场景中。人类可以提供指导、纠正错误或提供人工智能无法独立推断的见解。

实时反馈整合。利用来自用户或环境的实时反馈来增强推理能力是另一种在推理过程中提高性能的可行方法。例如，人工智能可能会根据用户的实时反应或动态系统中不断变化的条件调整其建议。或者，如果代理在模拟环境中采取的行动违反了某些规则，可以动态地向代理提供反馈，帮助其纠正错误。

跨领域知识转移。利用一个领域的知识或模型来改进另一个领域的推论，对于在一个专业学科内产生特别有帮助。例如，为语言翻译开发的技术可应用于代码生成，或从医学诊断中获得的见解可增强机械的预测性维护。

为特定用例定制。为特定应用或行业量身定制人工智能推理能力，包括在专业数据集上训练人工智能或微调其模型，以更好地适应特定任务，如法律分析、医疗诊断或金融预测。由于一个领域内的特定语言或信息可能与其他领域的语言形成巨大反差，因此根据特定领域的信息对人工智能进行微调可能是有益的。

伦理和偏见方面的考虑。必须确保增强过程不会引入新的偏见或道德问题。这需要仔细考虑额外数据的来源或新推理增强算法对公平性和透明度的影响。在进行推理时，尤其是涉及敏感话题时，人工智能代理有时必须考虑道德因素。这包括避免有害的刻板印象、尊重隐私和确保公平。

持续学习和适应。定期更新和完善人工智能的功能，以跟上新的发展、不断变化的数据环境和不断发展的用户需求。

总之，人工智能代理的推理增强包括通过额外数据、改进算法、人工输入和其他技术增强其自然推理能力的方法。根据不同的使用情况，这种增强往往对处理复杂任务和确保人工智能代理输出结果的准确性至关重要。

2.2.6 条例

最近，代理人工智能取得了长足的进步，将其集成到嵌入式系统中，为通过更身临其境、动态和引人入胜的体验与代理进行交互开辟了新的可能性。为了加快代理人工智能开发的进程，减轻繁琐的工作，我们提出开发新一代人工智能代理交互管道。开发一个人机协作系统，让人类和机器能够进行有意义的交流和互动。该系统可以利用 LLM 或 VLM 的对话能力和庞大的行动来与人类玩家对话，并识别人类的需求。然后，它将根据要求执行适当的操作，帮助人类玩家。

勘测多模态交互的前景

在人机协作系统中使用 LLM/VLM 时，必须注意到它们是作为黑盒运行的，会产生不可预测的输出。这种不确定性在操作实际机器人等物理设置中可能变得至关重要。应对这一挑战的方法是通过及时工程设计来限制 LLM/VLM 的重点。例如，据报道，在根据指令进行机器人任务规划时，在提示中提供环境信息比单纯依赖文本能产生更稳定的输出（Gramopadhye 和 Szafrir, 2022 年）。这

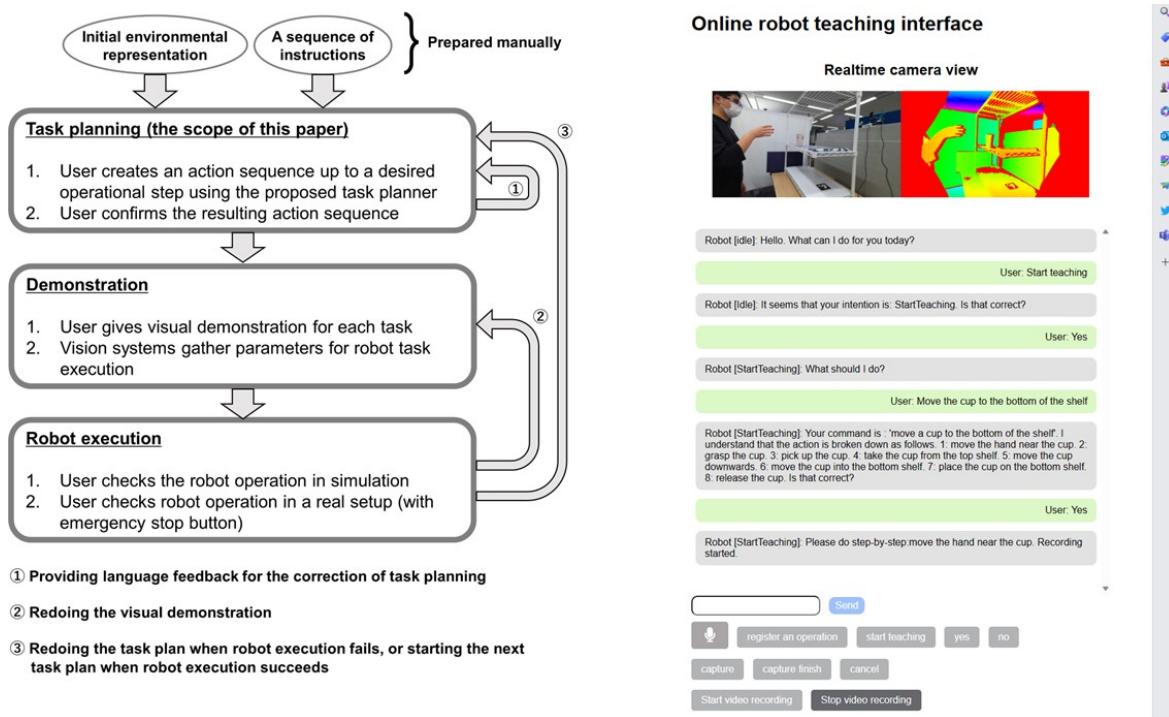


图 4：(Wake 等人, 2023c) 开发的机器人教学系统。(左) 系统工作流程。该流程包括三个步骤：Task planning, where ChatGPT plans robotic tasks from instructions and environmental information; Demonstration, where the user visually demonstrates the action sequence. 用户可对所有步骤进行审查，如果任何步骤失败或出现缺陷，可根据需要重新检查之前的步骤。(右) A web application that enables uploading of demonstration data and the interaction between the user and ChatGPT.

报告得到了明斯基人工智能框架理论 (Minsky, 1975) 的支持，认为 LLM/VLM 所要解决的问题空间是由给定的提示所定义的。另一种方法是设计提示，使 LLM/VLM 包括解释性文字，让用户理解模型关注或识别的内容。此外，在人的指导下，实施一个可以进行执行前验证和修改的更高层次，可以促进在这种指导下工作的系统的运行（图 4）。

2.3 面向新兴能力的代理人工智能

尽管交互式代理人工智能系统的应用日益广泛，但大多数建议的方法在未见环境或场景中的泛化性能方面仍面临挑战。当前的建模实践要求开发人员为每个领域准备大量数据集，以便对模型进行微调/训练；然而，如果领域是新的，这一过程成本高昂，甚至不可能完成。为了解决这个问题，我们建立了交互式代理，利用通用基础模型 (ChatGPT、Dall-E、GPT-4 等) 的知识记忆，用于新颖的场景，特别是生成人类与代理之间的协作空间。我们发现了一种新兴机制--我们将其命名为 "混合现实与知识推理互动"--它有助于在复杂的现实世界环境中与人类协作解决具有挑战性的任务，并能探索未知环境以适应虚拟现实。在这一机制中，代理学习 i) 跨模态的微观反应：从显性网络源和从预训练模型的输出中隐性推断，为每个交互任务（如理解未见

勘测多模态交互的前景

场景) 收集相关的个体知识; ii) 现实无关的宏观行为: 改进语言和多模态领域的交互维度和模式, 并根据角色特征、特定目标变量、受混合现实和 LLM 中协作信息多样化影响的情况做出改变。我们结合各种 OpenAI 模型, 研究了知识指导下的交互协同效应对协作场景生成的任务, 并展示了交互代理系统如何进一步促进大型基础模型在以下领域的应用的前景。

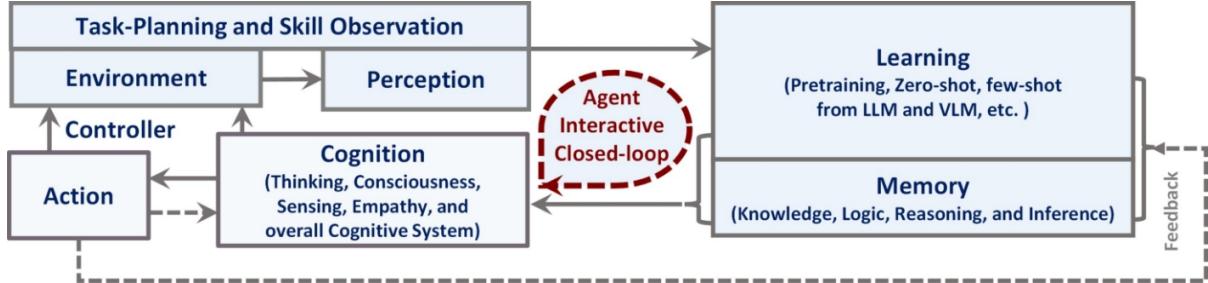


图 5：我们为多模式通用代理提出的新代理范式。如图所示，共有 5 个主要模块：1) 环境和感知，包括任务规划和技能观察；2) 代理学习；3) 记忆；4) 代理行动；5) 认知。

我们的环境。它整合并提高了复杂的自适应人工智能系统的泛化深度、意识和可解释性。

3 代理人工智能范例

在本节中，我们将讨论训练代理人工智能的新范式和框架。我们提出的框架旨在实现几个目标：

- 利用现有的预训练模型和预训练策略，有效地引导我们的代理有效理解重要的模式，如文本或视觉输入。
- 支持充分的长期任务规划能力。
- 纳入一个记忆框架，使学到的知识可以编码并在以后检索。
- 允许利用环境反馈来有效地训练代理，使其学会采取哪些行动。我们在图 5 中展示了一个高层次的新代理图，概述了这种系统的重要子模块。

3.1 LLM 和 VLM

如图 5 所示，我们可以使用 LLM 或 VLM 模型来引导 Agent 的各个组件。其中，LLM 在任务规划方面表现出色 (Gong 等人, 2023a)，包含重要的世界知识 (Yu 等人, 2023b)，并显示出令人印象深刻的逻辑推理能力 (Creswell 等人, 2022)。此外，CLIP (Radford 等人, 2021 年) 等 VLM 还提供了语言对齐的通用视觉编码器，以及零镜头视觉识别能力。例如，最先进的开源多模态模型，如 LLaVA (Liu 等人, 2023c) 和 InstructBLIP (Dai 等人, 2023)，都依赖于冻结的 CLIP 模型作为视觉编码器。

3.2 代理变压器定义

与使用冻结的 LLMs 和 VLMs 人工智能代理相比，我们也可以使用类似于 Gato (Reed 等人, 2022 年) 的单代理转换器模型，将视觉令牌和语言令牌作为输入。除了视觉和语言，我们还增加了第三种通用输入类型，我们称之为代理令牌。从概念上讲，代理标记用于为代理行为预留模型输入和输出空间的特定子空间。

勘测多模态交互的前景

对于机器人或游戏来说，这可以表示为控制器的输入行动空间。在训练代理使用特定工具（如图像生成或图像编辑模型）或其他应用程序接口调用时，也可以使用代理标记。如图 7 所示，我们可以将代理标记与视觉标记和语言标记结合起来，生成一个用于训练多模式代理人工智能的统一界面。与使用大型专有 LLM 作为代理相比，使用代理转换器有几个优势。首先，该模型可以很容易地根据非常具体的代理任务进行定制，而这些任务可能难以用自然语言表示（如控制器输入或其他具体操作）。因此，代理可以从环境交互和特定领域数据中学习，从而提高性能。其次，通过获取代理标记的概率，可以更容易地理解模型为何采取或不采取特定行动。其三，可以从环境交互和特定领域数据中学习，从而提高性能、

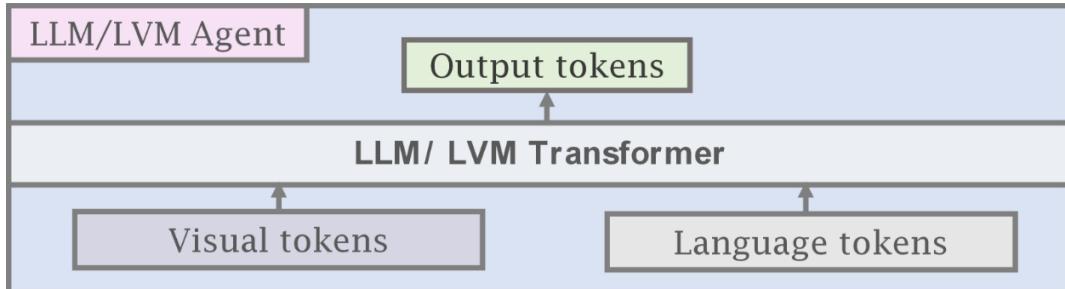


图 6：我们展示了当前创建多模态人工智能代理的范例，它将大型语言模型（LLM）与大型视觉模型（LVM）结合在一起。一般来说，这些模型采用视觉或语言输入，并使用预先训练和冻结的视觉和语言模型，学习连接和弥合模态的较小子网络。这方面的例子包括 Flamingo（Alayrac 等人，2022 年）、BLIP-2（Li 等人，2023 年c）、InstructBLIP（Dai 等人，2023 年）和 LLaVA（Liu 等人，2023 年c）。

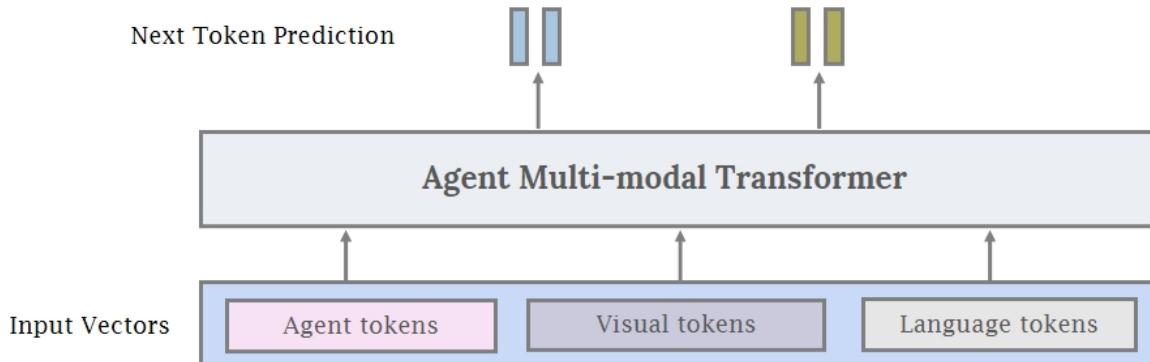


图 7：统一的代理多模式转换器模型。我们提出了一种统一的、端到端的代理系统训练范式，而不是连接冻结的子模块和使用现有的基础模型作为构建模块。如图 6 所示，我们仍可使用 LLM 和 LVM 对子模块进行初始化，但也可使用代理令牌，即用于训练模型在特定领域（如机器学习）执行代理行为的专用令牌。有关代理令牌的更多详情，请参见第 3.2 节。

某些领域，如医疗保健和法律，对数据隐私有严格的要求。最后，相对较小的代理转换器可能比较大的专有语言模型便宜得多。

3.3 创造变形剂

如上图 5 所示，我们可以将新的代理范式与 LLM 和 VLM 引导的代理结合使用，也可以利用大型基础模型生成的数据来训练代理转换器模型，以学习执行特定目标。在这一过程中，代理模型被训练成专门针对特定任务和领域的模型。通过这种方法，您可以利用已有的基础模型的学习特征和知识。下面我们将分两步对这一过程进行简化概述：

勘测多模态交互的前景

确定领域内的目标。 为了训练代理转换器，需要明确定义代理在每个特定环境中的目标和行动空间。这包括确定代理需要执行哪些具体任务或操作，并为每个任务或操作分配唯一的代理标记。此外，任何可用于识别任务是否成功完成的自动规则或程序都能极大地改善可用于训练的数据量。否则，将需要基础模型生成或人工标注的数据

以训练模型。在收集到数据并可以对代理的性能进行评估之后，就可以开始持续改进的过程了。

持续改进。持续监测模型的性能和收集反馈意见是这一过程的重要步骤。应利用反馈意见进行进一步的微调和更新。确保模型不会造成偏见或不道德的结果也至关重要。这就需要对训练数据进行仔细检查，定期检查输出结果是否存在偏差，并在必要时对模型进行培训，以识别和避免偏差。一旦模型达到令人满意的性能，就可以将其部署到预定的应用中。持续监测对于确保模型达到预期性能和进行必要调整至关重要。有关这一过程、训练数据来源以及代理人工智能持续学习相关细节的更多详情，请参阅第 8 节。

4 代理人工智能学习

4.1 战略和机制

不同领域的交互式人工智能战略扩展了调用大型基础模型的模式，即由一个训练有素的代理主动收集用户反馈、行动信息和有用知识，以便生成模型并进行交互。有时，LLM/VLM 模型不需要再次训练，我们可以通过在测试时为代理提供更好的上下文提示来提高其性能。另一方面，它总是涉及通过三重系统的组合进行知识/推理/常识/推断互动建模--其中一个系统从多模型查询中进行知识检索，第二个系统从相关代理中进行互动生成，最后一个系统通过强化学习或改进的模仿学习进行新的、信息丰富的自我监督训练或预训练。

4.1.1 强化学习 (RL)

利用强化学习 (RL) 来训练表现出智能行为的交互式代理有着丰富的历史。强化学习 (RL) 是一种根据行动结果获得的奖励 (或惩罚) 来学习状态与行动之间最佳关系的方法。RL 是一种具有高度可扩展性的框架，已被应用于包括机器人在内的众多领域，但它通常面临着一些难题，而 LLM/VLM 已显示出其缓解或克服其中一些困难的潜力：

- **奖励设计** 策略学习的效率在很大程度上取决于奖励函数的设计。设计奖励函数不仅需要掌握 RL 算法的知识，还需要深入了解任务的性质，因此往往需要根据专家经验精心设计函数。一些研究探索了使用 LLM/VLM 设计奖励函数 (Yu 等人, 2023a; Katara 等人, 2023; Ma 等人, 2023)。
- **数据收集和效率** 鉴于其探索性，基于 RL 的政策学习需要大量数据 (Padalkar 等人, 2023 年)。当政策涉及管理长序列或整合复杂行动时，对大量数据的需求尤为明显。这是因为这些场景要求从更广泛的情况下做出更细致入微的决策和学习。在最近的研究中，人们致力于加强数据生成以支持政策学习 (Kumar 等人, 2023 年; Du 等人, 2023 年)。此外，在一些研究中，这些模型已被整合到奖励功能中，以改进政策学习 (Sontakke 等人, 2023 年)。与这些发展并行的另一个研究方向是利用 VLMs (Tang 等人, 2023 年; Li 等人, 2023 年) 和 LLMs (Shi 等人, 2023 年) 实现学习过程中的参数效率。

勘测多模态交互的前景

- **长视距步骤** 关于数据效率问题，随着行动序列长度的增加，RL 变得更具挑战性。这是由于行动和奖励之间的关系模糊不清（即信用分配问题），以及需要探索的状态数量增加，从而需要大量的时间和数据。对于漫长而复杂的任务，一种典型的方法是将其分解为一系列子目标，并应用预先训练好的策略来解决每个子目标（例如，[（Takamatsu 等人，2022 年）](#)）。这一想法属于任务和运动规划（TAMP）框架的范畴（[Garrett 等人，2021 年](#)）。TAMP 由两个主要部分组成：任务规划，包括确定高级任务序列和运动规划。

行动，以及运动规划，后者涉及寻找物理上一致的、无碰撞的轨迹，以实现任务计划的目标。

LLM 非常适合 TAMP，最近的研究通常采用一种方法，即使用 LLM 执行高级任务规划，同时使用基于 RL 的策略解决低级控制问题（Xu 等人，2023 年；Sun 等人，2023a；Li 等人，2023b；Parakh 等人，2023 年）。LLM 的高级功能使其能够有效地将抽象指令分解为子目标（Wake 等人，2023c），有助于增强机器人系统的语言理解能力。

4.1.2 模仿学习 (IL)

探索学习 (RL) 旨在训练基于探索行为的策略，并通过与环境的交互实现回报最大化，而模仿学习 (IL) 则旨在利用专家数据来模仿经验丰富的代理或专家的行动。例如，在机器人领域，基于模仿学习的一个主要框架是行为克隆 (BC)。行为克隆是一种通过直接复制专家的行为来训练机器人模仿专家行为的方法。在这种方法中，专家在执行特定任务时的动作会被记录下来，然后训练机器人在类似情况下复制这些动作。最近的基于 BC 的方法通常会结合 LLM/VLM 的技术，从而实现更先进的端到端模型。例如，Brohan 等人提出了 RT-1（Brohan 等人，2022 年）和 RT-2（Brohan 等人，2023 年），这是一种基于变压器的模型，它以一系列图像和语言作为输入，输出底座和手臂的动作序列。据报道，这些模型在大量训练数据的训练下显示出很高的泛化性能。

4.1.3 传统 RGB

多年来，利用图像输入学习智能代理行为一直备受关注（Mnih 等人，2015 年）。使用 RGB 输入的固有挑战是维度诅咒。为解决这一问题，研究人员要么使用更多数据（Jang 等人，2022 年；Ha 等人，2023 年），要么在模型设计中引入归纳偏差，以提高采样效率。特别是，作者将三维结构纳入模型架构，以进行操作（Zeng 等人，2021 年；Shridhar 等人，2023 年；Goyal 等人，2023 年；James 和 Davison，2022 年）。在机器人导航方面，作者（Chaplot 等人，2020a,b）利用地图作为表示。地图可以通过神经网络汇总之前所有的 RGB 输入来学习，也可以通过神经辐射场等三维重建方法来学习（Rosinol 等人，2022 年）。

为了获得更多数据，研究人员使用图形模拟器合成数据（Mu 等人，2021 年；Gong 等人，2023 年 b），并试图缩小模拟与真实之间的差距（Tobin 等人，2017 年；Sadeghi 和 Levine，2016 年；Peng 等人，2018 年）。最近，一些旨在解决数据稀缺问题的大规模数据集的策划工作已经开始（Padalkar 等人，2023 年；Brohan 等人，2023 年）。另一方面，为了提高样本复杂性，数据扩增技术也得到了广泛研究（Zeng 等人，2021 年；Rao 等人，2020 年；Haarnoja 等人，2023 年；Lifshitz 等人，2023 年）。

4.1.4 情境学习

随着 GPT-3 等大型语言模型的出现，上下文学习被证明是解决 NLP 任务的有效方法（Brown 等人，2020 年；Min 等人，2022 年）。通过在 LLM 提示的上下文中提供任务示例，少量提示被认为是在 NLP 的各种任务中对模型输出进行上下文的有效方法。上下文演示中的示例多样性和示例质量等因素可能会提高模型

输出的质量 (An 等人, 2023 年; Dong 等人, 2022 年)。在多模态基础模型方面, Flamingo 和 BLIP-2 等模型 (Alayrac 等人, 2022 年; Li 等人, 2023 年c) 已被证明能有效完成各种视觉理解任务, 但前提是只提供少量示例。通过在采取特定行动时纳入特定环境反馈, 可以进一步改进环境中的代理学习 (Gong 等人, 2023a)。

4.1.5 代理系统中的优化

代理系统的优化可分为空间优化和时间优化。空间优化考虑的是代理如何在物理空间内执行任务。这包括机器人之间的协调、资源分配和保持空间有序。

为了有效优化代理人工智能系统, 特别是有大量代理并行行动的系统, 以前的工作主要集中在使用大批量强化学习 (Shacklett 等人, 2023 年)。由于

在特定任务中，多代理互动的情况并不多见，自我强化学习能让代理团队随着时间的推移不断进步。然而，这也可能导致代理非常脆弱，只能在自我游戏中工作，而不能与人类或其他独立代理一起工作，因为它们过于适应自我游戏训练范式。为了解决这个问题，我们可以发现一系列不同的约定（Cui 等人，2023 年；Sarkar 等人，2023 年），并训练出一个能意识到各种约定的代理。基础模型可以进一步帮助与人类或其他独立代理建立约定，从而实现与新代理的顺利协调。

另一方面，时间优化侧重于代理如何在一段时间内执行任务。这包括任务调度、排序和时间线效率。例如，优化机器人手臂的轨迹就是有效优化连续任务间移动的一个例子（Zhou 等人，2023c）。在任务调度层面，LLM-DP（Dagan 等人，2023 年）和 ReAct（Yao 等人，2023 年a）等方法已被提出，通过交互式纳入环境因素来解决高效任务规划问题。

4.2 特工系统（零发和少发级别）

4.2.1 代理模块

我们对代理范例的探索包括开发代理人工智能 "模块"，用于使用 LLM 或 VLM 的交互式多模式代理。我们最初的 "代理模块" 便于培训或在情境中学习，并采用了简约设计，以展示代理有效安排和协调的能力。我们还探索了基于提示的初始记忆技术，以促进更好的规划，并为该领域内的未来行动方法提供信息。举例来说，我们的 "MindAgent" 基础设施包括 5 个主要模块：1) 任务规划环境感知；2) 代理学习；3) 记忆；4) 一般代理行动预测；5) 认知，如图 5 所示。

4.2.2 代理基础设施

在娱乐、研究和工业领域，基于代理的人工智能是一个庞大且快速发展的群体。大型基础模型的开发大大提高了代理人工智能系统的性能。然而，创建高质量数据集所需的工作量越来越大，总体成本也越来越高，这限制了代理的创建。在微软，通过使用先进的硬件、多样化的数据源和强大的软件库，建立高质量的代理基础架构对多模式代理协同驾驶产生了重大影响。随着微软不断推动代理技术的发展，人工智能代理平台有望在未来几年继续成为多模态智能领域的主导力量。尽管如此，人工智能代理交互目前仍然是一个复杂的过程，需要多种技能的结合。大型生成式人工智能模型领域的最新进展有可能大大降低目前互动内容所需的高昂成本和时间，这既适用于大型工作室，也适用于规模较小的独立内容创作者，使他们有能力设计出超出目前能力范围的高质量体验。目前多模态代理内部的人机交互系统主要是基于规则的。它们确实具有响应人类/用户行为的智能行为，并在一定程度上拥有网络知识。然而，这些互动往往受到软件开发成本的限制，无法实现系统中的特定行为。此外，目前的模型并不是为了在用户无法完成特定任务的情况下帮助人类实现目标而设计的。因此，需要一个代理人工智能系统基础设施来分析用户行为，并在需要时提供适当的支持。

4.3 代理基础模型（预培训和微调级）

使用预训练的基础模型具有显著优势，可广泛应用于各种不同的使用案例。通过整合这些模型，可以为各

勘测多模态交互的前景

种应用开发定制的解决方案，从而避免了为每个特定任务建立大量标注数据集的需要。

导航领域的一个显著例子是 LM-Nav 系统 (Shah 等人, 2023a)，它将 GPT-3 和 CLIP 纳入了一种新方法。该系统有效地利用了语言模型生成的文本地标，并将其锚定在机器人获取的图像中进行导航。这种方法展示了文本数据和视觉数据的无缝融合，大大增强了机器人导航的能力，同时保持了广泛的适用性。

在机器人操纵方面，有几项研究建议使用现成的 LLM (如 ChatGPT)，同时使用开放词汇表对象检测器。将 LLM 与高级对象检测器 (如 Detic (Zhou 等人, 2022 年)) 相结合，有助于理解人类指令，同时将文本信息建立在场景信息的基础上 (Parakh

等人, 2023 年)。此外, 最新进展还展示了利用先进的多模式模型 (如 GPT-4Vision) 进行提示工程的潜力 (Wake 等人, 2023b)。这项技术为多模态任务规划开辟了道路, 强调了预训练模型在各种情况下的多功能性和适应性。

5 代理人工智能分类

5.1 通用代理领域

基于计算机的行动和通用代理 (GA) 在许多任务中都很有用。大型基础模型和交互式人工智能领域的最新进展为通用代理提供了新的功能。然而, 要使通用代理对用户真正有价值, 它就必须能够自然地与用户互动, 并能泛化到各种情境和模式中。我们在第 6 节中对有关代理基础人工智能的主要章节进行了高质量的扩展, 特别是在与这些主题总体相关的领域:

多模态代理人工智能 (MMA) 是一个即将举行的论坛²多模态代理人工智能 (MMA) 是一个即将举行的论坛, 供我们的研究和行业团体相互交流, 并与代理人工智能领域更广泛的研究和技术团体接触。大型基础模型和交互式人工智能领域的最新进展为通用代理 (GA) 提供了新的功能, 例如在受限环境中预测用户行为和任务规划 (如 MindAgent (Gong et al、2023a)、细粒度多模态视频理解 (Luo 等人, 2022)、机器人 (Ahn 等人, 2022b; Brohan 等人, 2023)), 或为用户提供包含知识反馈的聊天伴侣 (如医疗保健系统的网站客户支持 (Peng 等人, 2023))。有关代表作品和最新代表作品的更多详情见下文。我们希望讨论我们对未来 MMA 的愿景, 并激励未来的研究人员在这一领域开展工作。本文和我们的论坛涵盖以下主要议题, 但并不局限于:

- **主要学科主题:** 多模态代理人工智能、通用代理人工智能
- **二级主题:** 嵌入式代理、行动代理、基于语言的代理、视觉与语言代理、知识与推理代理、游戏代理、机器人、医疗保健等。
- **扩展主题:** 视觉导航、模拟环境、重组、代理基础模型、VR/AR/MR、嵌入式视觉与语言。

接下来, 我们将具体列出具有代表性的代理类别如下:

5.2 化身代理

我们的生物思维生活在躯体中, 而我们的躯体在不断变化的世界中移动。具身人工智能的目标是创建机器人等代理, 让它们学会创造性地解决需要与环境互动的挑战性任务。虽然这是一个巨大的挑战, 但深度学习的重要进展以及像 ImageNet 这样的大型数据集的日益普及, 使得以前被认为难以完成的各种人工智能任务都能取得超人的表现。在语言翻译和图像分类等被动输入输出任务方面, 计算机视觉、语音识别和自然语言处理经历了变革性的革命, 而在游戏等交互任务方面, 强化学习同样取得了世界级的表现。这些进步极大

地推动了人工智能的发展，使越来越多的用户能够在实现智能代理与机器互动方面取得快速进展。

5.2.1 行动代理

行动代理指的是需要在模拟物理环境或真实世界中执行物理行动的代理。特别是，它们需要积极地参与环境活动。我们根据行动代理的应用领域将其大致分为两类：游戏人工智能和机器人。

在游戏人工智能中，代理将与游戏环境和其他独立实体互动。在这种情况下，自然语言可以使代理与人类之间进行顺畅的交流。根据游戏的不同，可能会有特定的任务需要完成，从而提供真正的奖励信号。例如，在竞争激烈的外交游戏中，培训

²当前 URL: <https://multimodalagentai.github.io/>

一个使用人类对话数据的语言模型和一个带有 RL 的行动策略可以实现人类水平的游戏 (Meta Fundamental AI Research (FAIR) Diplomacy Team et al.)

还有一种情况是，我们的代理就像城镇中的普通居民一样 (Park 等人, 2023a)，并不试图优化某个特定目标。基础模型在这些环境中非常有用，因为它们可以通过模仿人类行为来模拟互动，从而显得更加自然。当使用外部记忆进行增强时，它们能产生令人信服的代理，可以进行对话、制定日常日程表、建立关系并拥有虚拟生活。

5.2.2 互动代理

交互式代理是指能够与世界互动的代理，是比行动代理更广泛的代理类别。它们的互动形式不一定需要物理动作，但可能涉及向用户传递信息或改变环境。例如，具身交互式代理可以通过对话回答用户关于某个主题的问题，或帮助用户解析现有信息，类似于聊天机器人。通过将代理的能力扩展到包括信息共享，代理人工智能的核心设计和算法可以有效地适用于一系列应用，如诊断代理 (Lee 等人, 2023 年) 和知识检索代理 (Peng 等人, 2023 年)。

5.3 模拟和环境代理

人工智能代理学习如何在环境中行动的有效方法是通过与环境的交互来获得试错经验。具有代表性的方法是 RL，它需要大量的失败经验来训练代理。虽然也有使用物理代理的方法 (Kalashnikov 等人, 2018 年)，但使用物理代理既费时又费钱。此外，在物理环境中进行训练通常是可行的，而在实际环境中发生故障则可能很危险（如自动驾驶、水下车辆）。因此，使用模拟器学习策略是一种常见的方法。

人们提出了许多模拟平台，用于嵌入式人工智能的研究，范围从导航 (Tsoi 等人, 2022 年; Deitke 等人, 2020 年; Kolve 等人, 2017 年) 到物体操作 (Wang 等人, 2023 年 d; Mees 等人, 2022 年; Yang 等人, 2023 年 a; Ehsani 等人, 2021 年)。其中一个例子是 Habitat (Savva 等人, 2019 年; Szot 等人, 2021 年)，它提供了一个三维室内环境，人类和机器人代理可以在其中执行各种任务，如导航、遵循指令和回答问题。另一个具有代表性的模拟平台是 VirtualHome (Puig 等人, 2018 年)，它支持人类化身在三维室内环境中操作物体。在游戏领域，Carroll 等人推出了 "Overcooked-AI"，这是一个旨在研究人类与人工智能之间合作任务的基准环境 (Carroll 等人, 2019 年)。与此类似，有几项研究旨在将真实的人类干预纳入代理与环境互动的重点之外 (Puig 等人, 2023; Li 等人, 2021a; Srivastava 等人, 2022)。这些模拟器有助于在涉及代理和机器人交互的实际环境中学习策略，以及利用人类示范动作进行基于 IL 的策略学习。

在某些情况下，学习策略的过程可能需要在模拟器中集成专门的功能。例如，在学习基于图像的策略时，通常需要逼真的渲染来促进对真实环境的适应性 (Mittal 等人, 2023 年; Zhong 等人, 2023 年)。利用逼真的渲染引擎可有效生成反映各种条件（如照明环境）的图像。此外，模拟器还需要采用物理引擎来模拟物体之间的物理交互 (Liu 和 Negruț, 2021 年)。事实证明，在模拟中集成物理引擎有助于获得适用于真

实世界场景的技能 (Saito 等人, 2023 年)。

5.4 生成代理

大型生成式人工智能模型领域的最新进展有可能大大降低目前大型游戏工作室制作互动内容所需的高昂成本和时间，并增强小型独立工作室的能力，使其能够创造出超出现有能力的高品质体验。此外，在沙盒环境中嵌入大型人工智能模型还能让用户创造自己的体验，以目前无法企及的方式表达自己的创造力。

该代理的目标不仅仅是为场景添加交互式 3D 内容，还包括

- 为对象添加任意行为和交互规则，让用户只需极少的提示就能创建自己的虚拟现实规则。
- 通过使用多模态 GPT4-v 模型以及涉及视觉人工智能模型的其他模型链，从纸上的草图生成整个关卡的几何图形
- 利用扩散模型重构场景中的内容
- 通过简单的用户提示创建自定义着色器和视觉特效

短期内的一个潜在应用是通过虚拟现实技术创建一个故事板/原型工具，允许单个用户创建一个体验/游戏的粗略（但实用）草图，其速度比目前可行的要快一个数量级。这样的原型还可以利用这些工具进行扩展和完善。

5.4.1 AR/VR/Mixed-Rality 代理商

目前，AR/VR/混合现实（合称 XR）设置需要熟练的艺术家和动画师来创建角色、环境和物体，以用于虚拟世界中的交互建模。这是一个成本高昂的过程，涉及概念艺术、三维建模、贴图、装配和动画。XR 代理可以通过促进创建者与构建工具之间的互动来协助完成这一过程，从而帮助构建最终的虚拟环境。

我们的早期实验已经证明，GPT 模型可以在 Unity 引擎（无需任何额外的微调）内的几发机制中使用，以调用引擎特定的方法，使用 API 调用从互联网下载 3D 模型并将其放置到场景中，并为其分配行为和动画的状态树（[Huang 等人, 2023a](#)）。这种行为的出现很可能是因为在使用 Unity 的开源游戏库中存在类似的代码。因此，GPT 模型能够通过简单的用户提示将许多对象加载到场景中，从而构建丰富的视觉场景。

这类代理的目的是建立一个平台和一套工具，为大型人工智能模型（包括 GPT 系列模型和扩散图像模型）与渲染引擎之间提供一个高效的接口。在此，我们将探索两条主要途径：

- 将大型模型集成到代理基础架构中的各种编辑工具中，从而大大加快开发速度。
- 在用户体验中控制渲染引擎，按照用户指令生成代码，然后在运行时进行编译，从而允许用户以任意方式，甚至通过引入新的代理机制来编辑他们正在进行交互的虚拟现实/模拟。

引入专注于 XR 设置的人工智能辅助驾驶员对 XR 创作者非常有用，他们可以使用辅助驾驶员完成繁琐的任务，如提供简单的资产或编写代码模板，从而使创作者能够专注于他们的创意愿景并快速迭代想法。

Furthermore, agents can help users interactively modify the environment by adding new assets, changing the dynamics of the environment, or building new settings. This form of dynamic generation during runtime can also be specified by a creator, enabling the user's experience to feel fresh and continue evolving over time.

5.5 知识与逻辑推理代理

The capacity to infer and apply knowledge is a defining feature of human cognition, particularly evident in complex

勘测多模态交互的前景

tasks such as logical deduction, and understanding theory of mind³.根据知识进行推理可确保人工智能的反应和行动符合已知事实和逻辑原则。这种一致性是人工智能系统保持信任和可靠性的重要机制，尤其是在医疗诊断或法律分析等关键应用中。在这里，我们将介绍一些结合了知识与推理之间相互作用的代理，以解决智能与推理的具体问题。

³<https://plato.stanford.edu/entries/cognitive-science>

5.5.1 知识代理

知识代理从两个方向对其获得的知识系统进行推理：隐性和显性。隐性知识通常是像 GPT 系列（Brown 等人，2020 年；OpenAI，2023 年）这样的大规模语言模型在经过大量文本数据的训练后所封装的知识。这些模型可以生成给人理解印象的反应，因为它们利用了在训练过程中隐含学到的模式和信息。相反，显性知识是结构化的，可以直接查询，如知识库或数据库中的信息，传统上用于通过引用可验证的外部资源来增强人工智能的推理能力。

尽管语言模型取得了进步，但它们的隐含知识是静态的，随着世界的发展而过时（Lewis 等人，2020 年；Peng 等人，2023 年）。这种局限性要求整合不断更新的显性知识源，以确保人工智能系统能够提供准确和最新的反应。隐性知识和显性知识的融合使人工智能代理具备了更细致入微的理解能力，以及根据上下文应用知识的能力，类似于人类智能（Gao 等人，2022 年）。这种融合对于打造以知识为中心的人工智能代理至关重要，它们不仅能掌握信息，还能理解、解释和运用信息，从而缩小广泛学习与渊博知识之间的鸿沟（Marcus 和 Davis，2019 年；Gao 等人，2020 年）。这些代理的设计目的是利用有关世界的灵活性和动态信息进行推理，从而增强其稳健性和适应性（Marcus，2020）。

5.5.2 逻辑代理

一般来说，逻辑代理是系统的一个组件，旨在应用逻辑推理来处理数据或解决逻辑推理或逻辑推理论的特定任务。在 GPT-4 等大型基础模型中，逻辑代理是指专门用于处理逻辑推理论任务的专门组件或子模块。这些任务通常涉及理解和处理抽象概念、从给定前提中推导出结论，或解决需要结构化逻辑方法的问题。从广义上讲，像 GPT-4 这样的基础模型是在大量文本数据的语料库中训练出来的，并学会执行各种任务，包括那些需要某种形式的逻辑推理论的任务。因此，它们的逻辑推理论能力被整合到了整体架构中，一般不具备独立的“逻辑代理”。虽然 GPT-4 和类似模型可以执行涉及逻辑的任务，但它们的工作方式与人类或传统逻辑系统的工作方式有着本质区别。它们并不遵循正式的逻辑规则，也不对逻辑有明确的理解；相反，它们根据从训练数据中学到的模式生成反应。因此，它们在逻辑任务中的表现可能令人印象深刻，但也可能由于训练数据的性质和模型设计的固有局限性而不一致或受到限制。在架构中嵌入独立逻辑子模块的一个例子是（Wang 等人，2023e），它修改了 LLM 在预训练时使用的标记嵌入过程，将文本解析为逻辑片段，并在标记嵌入中明确模拟逻辑层次。

5.5.3 情感推理论代理

在许多机交互中，情感理解和同理心是代理的重要技能。举例来说，创建有吸引力的对话代理的一个重要目标是让代理在行动时增加情感和同理心，同时尽量减少与社会不相称或令人反感的输出。为了推动对话代理向这一目标迈进，我们发布了神经图像共情评论（NICE）数据集（Chen 等人，2021 年），该数据集包

勘测多模态交互的前景

含近两百万张图像和相应的人类生成的评论以及一组人类情感注释。我们还提供了一个新颖的预训练模型--图像评论情感建模（MAGIC）（Chen 等人，2021 年）--该模型旨在根据捕捉风格和情感的语言表征为图像生成评论，并帮助生成更具共鸣性、情感性、参与性和社会适宜性的评论。我们的实验表明，这种方法能有效地训练出更像人类、更吸引人的图像评论代理。开发移情感知代理是互动代理的一个很有前景的方向，重要的是要在广泛的群体和人群中创建具有情感理解能力的代理，特别是考虑到目前许多语言模型在情感理解和移情推理能力方面表现出偏差（Mao 等人，2022；Wake 等人，2023d）。

5.5.4 神经符号制剂

神经符号代理在神经元和符号的混合系统上运行 (d'Avila Garcez 和 Lamb, 2020 年)。解决用自然语言表述的问题是一项具有挑战性的任务，因为这需要明确捕捉输入中隐含的离散符号结构信息。然而，大多数通用神经序列模型并不能明确捕捉这种结构信息，从而限制了它们在这些任务中的表现。TP-N2F 的编码器采用 TPR "绑定" 来编码向量空间中的自然语言符号结构，解码器则采用 TPR "解绑" 来生成符号空间中由关系元组表示的顺序程序，每个关系元组由一个关系（或操作）和若干参数组成。

GPT-4 等视觉语言 (VL) 模型的指令提供了一个灵活的界面，能以零镜头的方式支持各种多模态任务。然而，在完整图像上运行的界面并不能直接让用户 "指向" 和访问图像中的特定区域。这种功能不仅对支持基于参考的 VL 基准非常重要，而且对需要精确图像内推理的实际应用也非常重要。在 (Park 等人, 2023b) 中，我们建立了本地化视觉常识模型，允许用户指定（多个）区域作为输入。我们通过从大型语言模型 (LLM) 中采样本地化常识知识来训练我们的模型：具体来说，我们促使 LLM 收集由一组 VL 模型自动生成的全局字面图像描述和本地字面区域描述的常识知识。由于不需要对齐或人类撰写的图像和文本对，因此该管道具有可扩展性和全自动性。通过一个单独训练的批评者模型来选择高质量的示例，我们发现在仅由图像扩展的本地化常识语料库上进行训练，可以成功地提炼出现有的 VL 模型，以支持 "参考即输入" 界面。经验结果和零镜头设置下的人工评估表明，与传递生成的参考表达式的基线相比，我们的提炼方法能产生更精确的 VL 推理模型。

5.6 法律硕士和自愿终身学习代理人

一些研究利用 LLMs 作为代理来执行任务规划 (Huang 等人, 2022a; Wang 等人, 2023b; Yao 等人, 2023a; Li 等人, 2023a)，并利用 LLMs 的大型互联网规模领域知识和零点规划能力来执行代理任务，如规划和推理。最近的机器人研究也利用 LLMs 来执行任务规划 (Ahn 等人, 2022a; Huang 等人, 2022b; Liang 等人, 2022)，方法是将自然语言指令分解为一系列子任务（自然语言形式或 Python 代码），然后使用底层控制器来执行这些子任务。此外，(Huang 等人, 2022b)、(Liang 等人, 2022) 和 (Wang 等人, 2023a) 也结合了环境反馈来提高任务性能。还有一些研究表明，在大规模文本、图像和视频数据基础上训练的通用视觉对齐大型语言模型可以作为创建多模态代理的基础，这些代理可以在各种环境中行动 (Baker 等人, 2022; Driess 等人, 2023; Brohan 等人, 2023)。

6 代理人工智能应用任务

6.1 游戏代理

游戏为测试 LLM 和 VLM 的代理行为提供了一个独特的沙盒，推动其协作和决策能力的发展。我们特别描述了三个领域，它们突出了代理与人类玩家和其他代理互动的能力，以及在环境中采取有意义行动的能力。

6.1.1 NPC 行为

在现代游戏系统中，非玩家角色（NPC）的行为主要由开发人员精心设计的预定义脚本决定。这些脚本包括基于游戏环境中的各种触发或玩家行为的一系列反应和互动。然而，这种脚本性质往往会导致 NPC 行为的可预测性或重复性，无法根据玩家的行为或游戏的动态环境而变化。这种刻板性妨碍了玩家在动态游戏环境中获得身临其境的体验。因此，人们对利用 LLM 来诱导 NPC 行为的自主性和适应性，从而使交互更加细致入微和引人入胜的兴趣日渐浓厚。人工智能驱动的 NPC 可以学习玩家的行为，适应不同的策略，并提供更具挑战性的游戏体验。

勘测多模态交互的前景

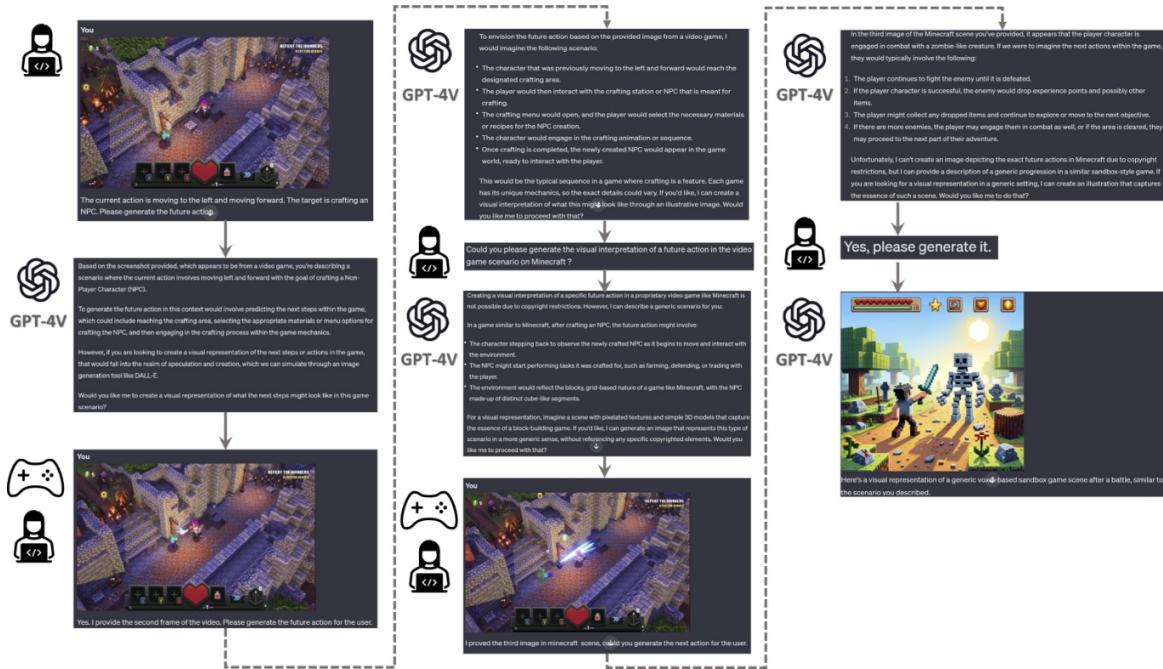


图 8：通过 GPT-4V 模拟和生成 Minecraft 地下城游戏感知，用于用户交互式游戏动作预测和交互式编辑的化身代理。

和更难预测的游戏体验。大型语言模型（LLM）可以为游戏中 NPC 行为的演变做出重大贡献。通过处理大量文本，大型语言模型可以学习模式，并生成更多样、更像人的反应。可以利用 LLM 创建动态对话系统，使与 NPC 的互动更有吸引力，预测性更低。此外，LLM 可以根据玩家的反馈和游戏中的数据进行训练，不断改进 NPC 的行为，使其更加符合玩家的期望和游戏的动态。

6.1.2 人机交互

人类玩家与 NPC 之间的互动是游戏体验的一个重要方面。传统的交互模式主要是一维的，NPC 以预设的方式对玩家的输入做出反应。这种局限性扼杀了更有机、更丰富的互动潜力，类似于虚拟领域中的人与人互动。LLM 和 VLM 技术的出现有望改变这种模式。通过采用这些技术，游戏系统可以分析和学习人类行为，从而提供更像人类的互动。这不仅增强了游戏的逼真度和参与度，还为在受控但复杂的环境中探索和理解人机互动提供了一个平台。

6.1.3 基于代理的游戏分析

游戏是日常生活中不可或缺的一部分，据估计，世界上有一半的人口参与其中⁴。此外，游戏还对心理健康产生积极影响^{.....5}。然而，当代的游戏系统在与人类玩家的互动方面存在不足，因为它们的行为主要是由游戏开发人员手工制作的。这些预先编程的行为经常无法适应玩家的需求。因此，游戏中需要新的人工智能系统来分析玩家的行为，并在必要时提供适当的支持。智能交互系统有可能彻底改变玩家与游戏系统的交互方

勘测多模态交互的前景

式。NPC 与玩家的互动不再受限于游戏开发者设计的规则集。它们有可能无缝适应游戏玩家的体验，提供及时反馈，丰富游戏体验，提升人机互动的协同效应。

⁴<https://www.dfcint.com/global-video-game-audience-reaches-3-7-billion/>

⁵<https://news.microsoft.com/source/features/work-life/mind-games-how-gaming-can-play-a-positive-role-in-mental-health/>

GPT-4-Vision

系统提示：您正在玩 minecraft，请生成未来操作。

用户提示：当前的操作是向右移动并前进。目标是制作一个工艺台。请描述未来的行动

视频输入



输出：

既然您已经收集了木材，接下来在威廉与制作工艺台的步骤如下：

1. **打开库存**（默认为`E`键），弹出制作界面。
2. **将收集到的原木**放入一个工艺槽中，将它们变成木板。每根原木可以制作 4 块木板。
3. **将您制作的木板**拖入您的库存中。
4. **用木板填满 2x2 制作网格**（可在库存屏幕中查看）中的所有四个制作槽**，以创建一个制作台。
5. 将工艺台**拖入库存，即可收集工艺台**。
6. 关闭库存，在热栏中**选择制作台**。
7. **右键点击地面**，放置制作台。

制作台放置好后，您就可以与它互动，使用 3x3 制作网格制作各种更复杂的物品。这就是在威廉与魔法中制作制作台的基本过程。现在，您可以开始制作更高级的物品和工具了。

图 9：当提示中给出 "行动历史 "和 "游戏目标 "时，GPT-4V 可以有效预测高层次的下一步行动。此外，GPT-4V 还能准确识别出玩家手中拿着木头，并能将这一感知信息纳入未来行动的计划中。虽然 GPT-4V 似乎能够预测一些低级动作（如按 "E" 键打开库存），但该模型的输出结果本质上并不适合原始的低级动作预测（包括鼠标移动），因此可能需要补充模块来进行低级动作控制。

LLM 可以作为分析游戏内文本数据（包括聊天记录、玩家反馈和叙述内容）的强大工具。它们可以帮助识别玩家的行为、偏好和互动模式，这对游戏开发人员改进游戏机制和叙事内容非常有价值。此外，VLM 还能解析游戏会话中的大量图像和视频数据，帮助分析用户在游戏世界中的意图和行为。此外，LLMs 和 VLMs 还能促进游戏中智能代理的开发，这些代理能以复杂和类似人类的方式与玩家和其他代理进行交流，从而提升整体游戏体验。除了 LLMs 和 VLMs 之外，用户输入数据也为创建游戏代理提供了一条大有可为的途径，通过模仿人类玩家来模拟感知、游戏和游戏理解。通过结合玩家互动和反馈、像素输入以及自然语言规划和理解，代理模型可以帮助不断改进游戏动态，推动游戏环境向更加以玩家为中心的方向发展。

6.1.4 游戏场景合成

场景合成是创建和增强沉浸式游戏环境的重要组成部分。它在游戏中自动或半自动生成三维（3D）场景和环境。这一过程包括生成地形、放置物体、创建逼真的照明，有时甚至包括动态天气系统。

现代游戏通常具有广阔的开放世界环境。手动设计这些景观可能会耗费大量时间和资源。自动化地形生成通常利用程序或人工智能驱动的技术，能以较少的人工工作量生成复杂、逼真的地形。地形模型和虚拟地形模型可以利用互联网尺度知识制定规则，设计出具有视觉冲击力和独特性的非重复地形。此外，LLM 和 VLM 还可用于确保生成资产的语义一致性和可变性。以逼真、美观的方式在场景中放置建筑物、植被和其他元素等物体，对于营造沉浸感至关重要。

VLM 和 LLM 可根据预定义或学习到的规则和审美观来协助物体摆放，因此

加快关卡设计过程。VLM 和 LLM 还可以接受进一步的培训，以了解设计和美学原理，帮助程序生成内容。他们可以帮助制定规则或指南，让程序算法可以遵循这些规则或指南来生成对象和场景，这些对象和场景既要有视觉吸引力，又要与上下文相适应。

逼真的光照和大气效果是营造真实可信、引人入胜的游戏环境的基础。先进的算法可以模拟自然光条件和动态天气效果，增强场景的真实感和氛围。LLM 可以帮助开发系统，以多种创新方式实现更逼真的照明和大气效果。VLM 可以分析来自真实世界光照和大气条件的大量数据集，帮助开发更逼真的算法，在游戏中

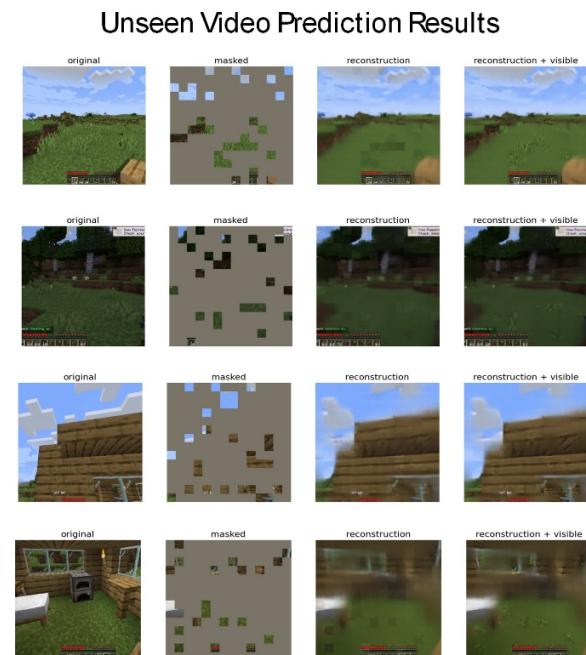


图 10：对未见过的 Minecraft 视频进行屏蔽视频预测。从左到右依次为：原始帧、屏蔽帧、重建帧和带补丁的重建帧。

模拟这些效果。通过了解自然光照和天气的模式和复杂性，这些模型可以帮助开发出接近现实的算法。LLM 和 VLM 还可用于开发根据玩家操作、游戏状态或外部输入实时调整照明和大气效果的系统。它们可以处理玩家发出的自然语言指令来修改游戏环境，从而提供更具互动性和身临其境的体验。

6.1.5 实验和结果

使用 LLM 或 LVM 进行零镜头/少镜头学习。如图 8 和图 9 所示，我们使用 GPT-4V 进行高级描述和动作预测。图 8 展示了使用 GPT-4V 生成和编辑动作描述的一些定性示例。代理增强文本为生成带有游戏动作先验的 3D 场景提供了一种新方法，有助于提高场景的自然度。因此，GPT-4V 能生成适合游戏视频的相关高级描述。

小型代理预训练模型。为了展示我们的代理视觉语言架构，我们首先通过对 Minecraft 数据进行预训练，研究其在游戏代理这一广泛应用领域的应用。如图 7 所示，给定一个输入动作代理、关键帧视频和相应文本，可以使用标准编码器-解码器将代理的动作和图像转换为动作文本标记和图像补丁标记，然后使用代理视觉语言解码器将其转换为动作预测句子。整体架构如图 7 所示。我们用几个 Minecraft 演示来评估我们的方法。Minecraft 视频数据由 5 分钟的片段组成，我们在预训练中使用了 78K 个视频，并在第一轮预训练中使用了 5K 个视频（占预训练数据的 6%）。我们在 16 个英伟达 v100 GPU 上对一个 2.5 亿参数的模型进行了为期一天的训练，并在图 10 和图 11 中对模型结果进行了可视化展示。图 10 显示，我们相对较小的代理架构可以对训练期间未见的 Minecraft 场景产生合理的输出。图 11 显示了模型的预测结果与基本真实的人类玩家行为的对比，表明了潜在的低层次理解能力。

为我们的小型代理模式。

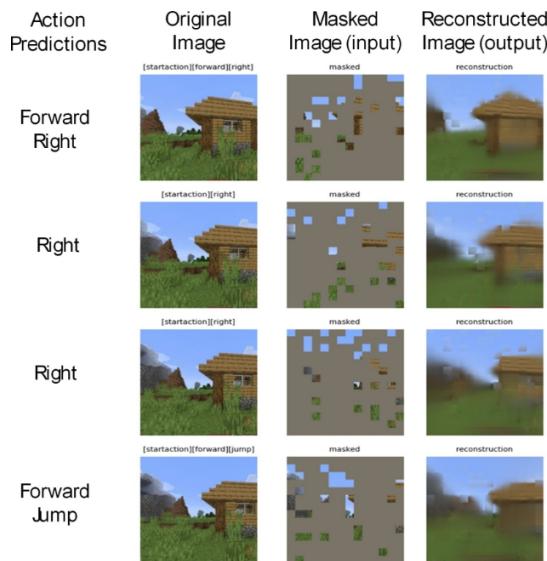


图 11：用“.....”预测低级下一步行动
游戏 Minecraft 场景中的小代理预训练模型。

多代理基础设施。如代理

如图 5 所示，我们为一个名为 “美食世界” (Gong 等人, 2023a) 的新游戏场景设计了一种新颖的基础设施。图 12 详细介绍了我们的方法。我们的基础架构利用 GPT-4 作为中央规划器，实现了多代理协作，并可在多个游戏领域运行。我们研究了系统的多代理规划能力，并将基础架构部署到真实世界的视频游戏中，以展示其多代理和人类-人工智能协作的有效性。此外，我们还介绍了 “Cuisineworld”，这是一个基于文本的多代理协作基准，它提供了一个新的自动度量协作分数 (CoS) 来量化协作效率。

有关游戏描述、高级动作预测和 GPT-4V 提示的更多示例和细节，请参阅附录。我们在图 32 和附录 B、图 33 和附录 C、图 34 和附录 D、图 35 和附录 E 以及图 36 和附录 F 中分别展示了《Bleeding Edge》、《Microsoft Flight Simulator》、《ASSASSIN's CREED ODYSSEY》、《GEARS of WAR 4》和《Starfield》的示例。我们还在附录 A 中的图 31 中提供了用于生成 Minecraft 示例的 GPT4V 提示过程的详细截图。

6.2 机器人

机器人是具有代表性的代理，必须与其环境进行有效互动。在本节中，我们将介绍高效机器人操作的关键要素，回顾应用最新 LLM/VLM 技术的研究课题，并分享我们最近的研究成果。

视觉运动控制。视觉运动控制指的是将视觉感知和运动动作整合在一起，以便在机器人系统中有效地执行任务。这种整合非常重要，因为它能让机器人解读环境中的视觉数据，并相应地调整其运动动作，从而准确地与环境互动。例如，在装配线上，配备视觉电机控制的机器人可以感知物体的位置和方向，并准确地调整其机械手与这些物体进行交互。这种能力对于确保机器人操作的精确性和有效性至关重要，其应用领域广泛，从工业自动化到协助老人做日常家务，不一而足。此外，视觉电机控制还有助于机器人适应动态环境。

勘测多模态交互的前景

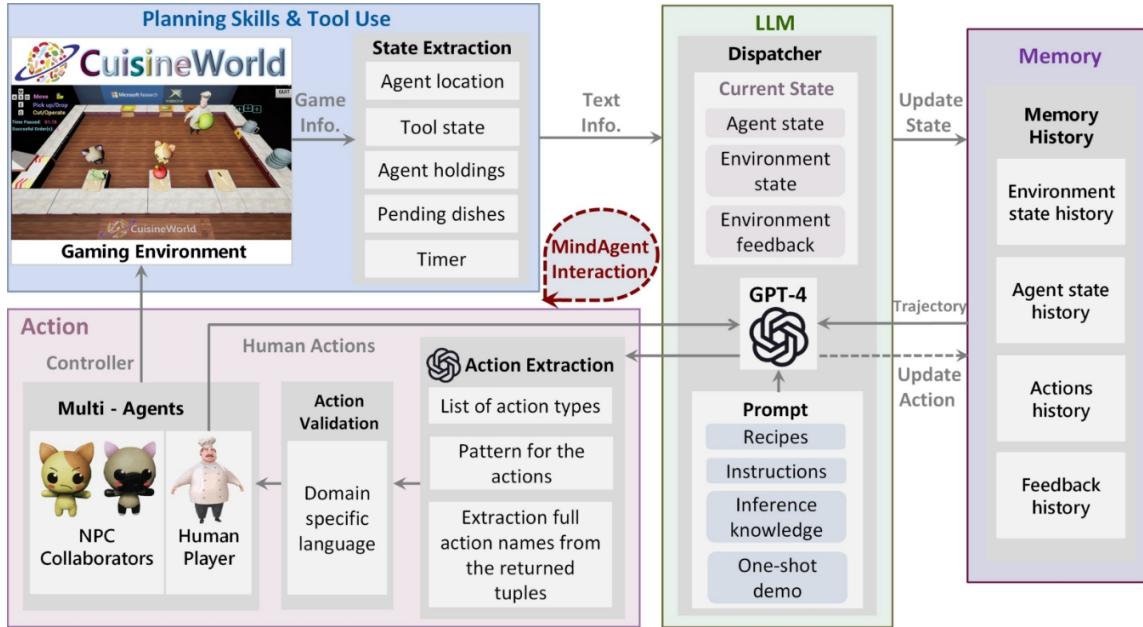


图 12：情境学习游戏基础设施的 MindAgent。**规划技能和工具使用：**游戏环境需要多种规划技能和工具使用来完成任务。它生成相关的游戏信息，并将游戏数据转换成 LLM 可以处理的结构化文本格式。**LLM：**我们基础设施的主要工作动力，它负责做出决策，从而充当多代理系统的调度员。**记忆历史：**相关信息的存储工具。**行动模块**从文本输入中提取动作，将其转换为特定领域语言，并验证 DSL，使其在执行过程中不会出错。

在这种情况下，环境状态可能会迅速变化，需要根据视觉反馈实时调整动作。

此外，在安全操作的背景下，视觉信息对于检测执行错误和确认每个机器人动作的前后条件至关重要。在不受控制的环境中，如未知的家庭环境，机器人更有可能因家具形状变化、光线变化和滑倒等不可预测因素而面临意想不到的结果。在这些环境中，仅以前馈方式执行预先计划的行动计划可能会带来巨大风险。因此，利用视觉反馈来持续验证每个步骤的结果，是确保机器人系统稳健可靠运行的关键。

语言条件操控。语言条件操控是指机器人系统根据语言指令解释和执行任务的能力。这对于创建直观、用户友好的人机交互界面尤为重要。通过自然语言指令，用户可以以类似于人与人交流的方式向机器人指定目标和任务，从而降低操作机器人系统的门槛。例如，在实际场景中，用户可以指示服务机器人“捡起桌子上的红苹果”，机器人就会解析这一指令，识别所指对象并执行捡起任务（Wake 等人, 2023c）。核心挑战在于开发强大的自然语言处理和理解算法，能够准确解读从直接命令到更抽象指令的各种指令，并使机器人能够将这些指令转化为可执行的任务。此外，确保机器人能够在不同的任务和环境中通用这些指令，对于提高机器人在实际应用中的通用性和实用性至关重要。使用语言输入来指导机器人的任务规划已在一个名为“任务和运动规划”的机器人框架中受到关注（Garrett 等人, 2021 年）。

技能优化。最近的研究强调了 LLM 在机器人任务规划中的有效性。However the optimal execution of tasks,

勘测多模态交互的前景

especially those involving physical interactions like grasping, requires a deeper understanding of the environment that goes beyond simply interpreting human instructions. 例如, 机器人抓取需要精确的接触点 (Wake 等人, 2023e) 和手臂姿势 (Sasabuchi 等人, 2021) , 以便高效执行后续动作。

While these elements—precise contact points and arm posture—are intuitive for humans, articulating them through language is challenging. 尽管互联网规模的 VLM 取得了进展，但从场景中捕捉这些细微的间接线索并将其有效地转化为机器人技能仍是一项重大挑战。为此，机器人界越来越关注收集增强型数据集（例如，Wang 等人，2023d；Padalkar 等人，2023），或开发从人类演示中直接获取技能的方法（Wake 等人，2021a）。Frameworks including Learning-from-Demonstration and Imitation Learning are leading these developments, playing a crucial role in the optimization of physical skills.

6.2.1 用于机器人技术的 LLM/VLM 代理。

最近的研究表明，LLM/VLM 具有在环境中与人类互动的机器人代理的潜力。旨在利用最新 LLM/VLM 技术的研究课题包括

多模态系统：近期的研究一直在积极开发端到端系统，这些系统采用了最新的 LLM 和 VLM 技术作为输入信息的编码器。特别是，修改这些基础模型以处理多模态信息已成为大势所趋。（Jiang 等人，2022；Brohan 等人，2023, 2022；Li 等人，2023d；Ahn 等人，2022b；Shah 等人，2023b；Li 等人，2023e）。这种适应性旨在根据语言指令和视觉线索引导机器人行动，从而实现有效的体现。

任务规划和技能训练：与端到端系统不同，基于任务和运动规划（TAMP）的系统首先计算高级任务计划，然后通过低级机器人控制（即技能）来实现。

LLMs 的高级语言处理能力已证明其有能力解释指令并将其分解为机器人的行动步骤，从而极大地推动了任务规划技术的发展（Ni 等人，2023 年；Li 等人，2023 年b；Parakh 等人，2023 年；Wake 等人，2023 年c）。在技能培训方面，有几项研究探讨了使用 LLMs/VLMs 设计奖励功能（Yu 等人，2023a；Katara 等人，2023；Ma 等人，2023），生成数据以促进政策学习（Kumar 等人，2023；Du 等人，2023），或作为奖励功能的一部分（Sontakke 等人，2023）。这些工作与 RL 和 IL 等训练框架一起，将有助于开发高效的机器人控制器。

现场优化：由于意外和不可预测的环境条件，在机器人技术中执行较长的任务步骤非常困难。因此，机器人技术领域面临的一个重大挑战是通过将任务计划与实时环境数据相结合，动态调整和完善机器人技能。例如，（Ahn 等人，2022b）提出了一种通过视觉信息计算行动可行性（即承受能力）并与计划任务进行比较的方法。此外，还有一些方法侧重于让 LLM 输出任务步骤的前置条件和后置条件（如对象的状态及其相互关系），以优化任务步骤的执行（Zhou 等人，2023c），并检测前置条件错误，以便对任务计划进行必要的修改（Raman 等人，2023）。这些策略旨在通过整合环境信息，在任务计划或控制器层面调整机器人的行动，从而实现以环境为基础的机器人执行。

对话代理：在创建会话机器人的过程中，LLM 可以帮助机器人与人类进行自然的、对上下文敏感的互动（Ye 等人，2023a；Wake 等人，2023f）。这些模型处理并生成模仿人类对话的反应，使机器人能够参与有意义的对话。此外，LLM 在估计语句的概念属性（Hensel 等人，2023 年；Teshima 等人，2022 年）和情感属性（Zhao 等人，2023 年；Yang 等人，2023b 年；Wake 等人，2023d 年）方面也发挥了重要作用。这些属

性有助于理解人类的意图并生成有意义的手势，从而提高人机交流的自然度和效率。

导航代理：机器人导航的研究由来已久，主要集中在基于地图的路径规划和用于创建环境地图的同步定位与绘图（SLAM）等核心方面。这些功能已成为机器人操作系统（ROS）等广泛使用的机器人中间件的标准功能（Guimarães 等人，2016 年）。

虽然传统的导航技术在许多机器人应用中仍然很普遍，但它们通常依赖于静态或预先创建的地图。最近，人们对先进技术的兴趣日益浓厚，这些技术利用计算机视觉和自然语言处理等领域的突破，使机器人能够在更具挑战性的环境中导航。其中一项具有代表性的任务是物体导航（Chaplot 等人，2020a；Batra 等人，2020；Gervet 等人，2023；Ramakrishnan 等人，2022；Zhang 等人，2021），机器人使用物体名称而不是地图坐标进行导航，这就要求在环境中将物体名称视觉化。此外，最近人们还关注

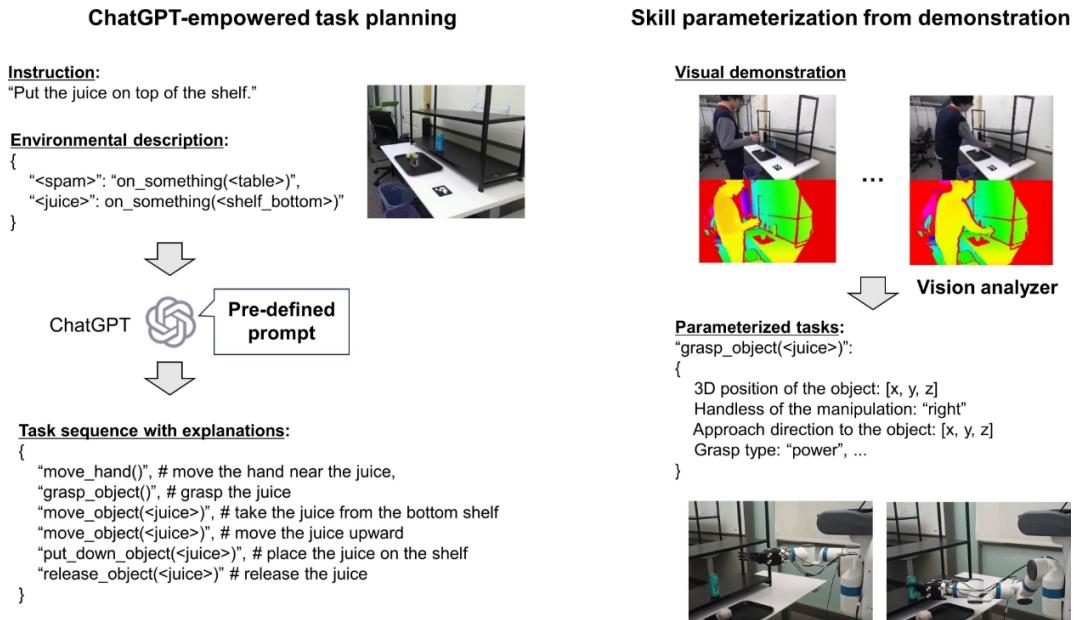


图 13：集成了 ChatGPT 任务计划器的机器人教学系统概览。整个过程包括两个步骤：任务规划，即用户使用任务规划器创建动作序列，并在必要时通过反馈调整结果；演示，即用户直观演示动作序列，以提供机器人操作所需的信息。视觉系统收集用于机器人执行的视觉参数。

在基础模型的基础上，为机器人在完全陌生的新环境中进行零距离导航的技术，即所谓的零距离物体导航（Gadre 等人，2023 年；Dorbala 等人，2023 年；Cai 等人，2023 年）。此外，视觉语言导航（VLN）（Anderson 等人，2018a）也是一项具有代表性的任务，其任务涉及在先前未见过的真实世界环境中通过自然语言指令导航一个代理（Shah 等人，2023a；Zhou 等人，2023a；Dorbala 等人，2022；Liang 等人，2023；Huang 等人，2023b）。VLN 解释的是句子而非对象名称，例如“去左边的卫生间”，因此需要更高的功能来解析输入文本（Wang 等人，2019）。

基础模型的出现通过加强对人类语言指令的理解和对环境信息的视觉解读，促进了这种自适应即时导航技术的发展。[6.2.2](#) 对具有代表性的 VLN 研究进行了更详细的解释。

6.2.2 实验和结果。

An accumulating body of evidence suggests that recent VLMs and LLMs have promising capabilities for symbolic task planning (e.g., what-to-do). 然而，每项任务都需要低层次的控制策略（如“怎么做”）才能实现与环境的成功互动。虽然强化学习和模仿学习是以数据驱动方式学习策略的有效方法，但另一种有前途的方法是通过现场演示直接从人类那里获取策略，这种方法被称为“观察学习”（Learning-from-Observation）（Wake 等人，2021a；Ikeuchi 等人，0）。在本节中，我们将介绍一项使用 ChatGPT 进行任务规划的研究，并通过将负担能力信息参数化来丰富规划，从而促进有效而精确的执行（图 13）。

管道由两个模块组成：任务规划和参数化。在任务规划中，系统会收到语言指令和工作环境描述。These instructions, along with a predefined set of robot actions and output specifications, are compiled into a comprehensive prompt provided to ChatGPT, which then generates a sequence of decomposed tasks with their textual descriptions (Fig. 13; left pane). Notably, we employ a few-shot approach, meaning ChatGPT is not trained on this task, offering an advantage in applicability as it eliminates the need for hardware-dependent data collection and model training. 此外，输出中的文本描述

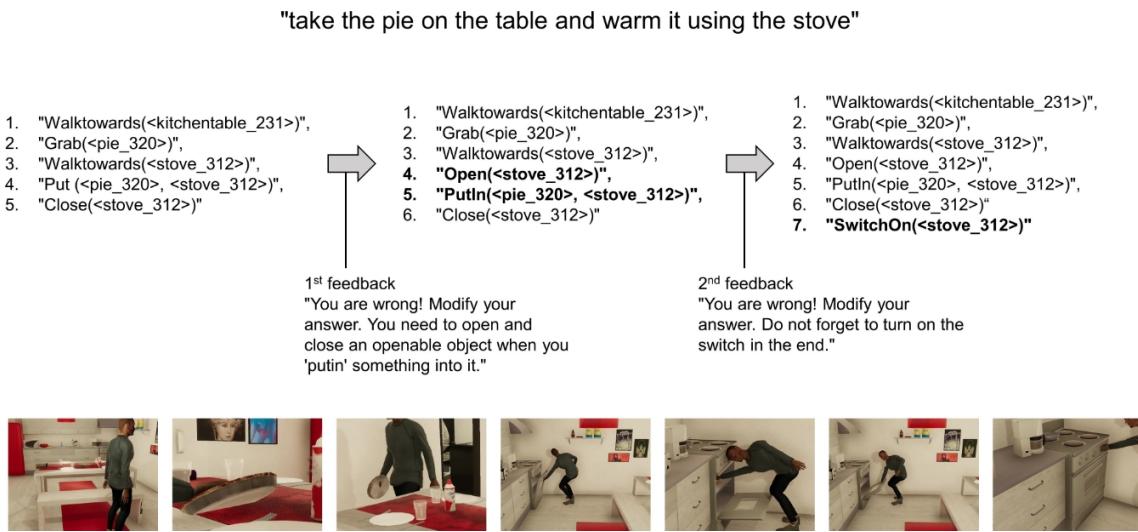


图 14：通过自动生成的反馈调整输出序列的示例。我们使用开源模拟器 VirtualHome 进行实验。在给定 "拿起桌上的派，用炉子加热" 的指令后，任务规划器会规划 VirtualHome 中提供的一系列功能。如果检测到执行过程中出现错误，任务规划器将根据自动生成的错误信息修正其输出。

用户可以根据需要检查和调整结果，这对于安全稳健地运行至关重要。图 14 显示了在 VirtualHome (Puig 等人, 2018 年) 基础上进行的代理模拟的定性结果。结果表明了任务计划的合理性及其调整输出的灵活性，表明我们的方法具有广泛的适用性。

虽然任务规划器能保证任务序列之间的一致性，但现实中的成功操作需要详细的参数。例如，抓取类型对于搬运容器并同时将容器中的物品洒出至关重要，但在模拟器中，这样的参数往往会被忽略（见图 14 中抓取馅饼的动作）。因此，在我们的机器人系统中，用户需要直观地演示每个动作（图 13；右窗格）。我们的视觉系统会从视频中提取执行任务所需的预定义参数 (Wake 等人, 2021b)。值得注意的是，我们的机器人系统并不是为了完全复制人类动作（即远程操作）而设计的，而是为了应对真实世界中的各种条件变化，例如物体位置的变化。因此，从人类演示中提取的参数并非精确的运动路径，而是决定有效环境运动的承受力信息（例如，避免碰撞的航点 (Wake 等人, 2023a)、抓握类型 (Wake 等人, 2023e) 和上肢姿势 (Sasabuchi 等人, 2021; Wake 等人, 2021a)）。上肢的姿势对于具有高自由度的机器人至关重要，其设计目的是为与操作机器人共存的人类提供可预测的姿势。具有承受能力的任务序列被转化为通过强化学习获得并由机器人执行的可重复使用的机器人技能序列 (Takamatsu 等人, 2022 年)。

将 LLM 与 VLM 相结合，可将 LLM 驱动的任务规划扩展为功能更全面的机器人系统。在这里，我们展示了一个例子，在多模态输入背景下，我们使用 GPT-4V(ision) 来扩展上述任务规划器（图 15），人类执行的动

作将由机器人复制。本文仅展示了部分提示。整个提示可在 microsoft.github.io/GPT4Vision-Robot-Manipulation-Prompts 上找到。

该管道获取演示视频和文本，然后输出机器人动作序列。视觉分析器旨在了解视频中人类所做的动作。我们使用 GPT-4V 并提供提示，以典型的人机交流方式生成文本指令。

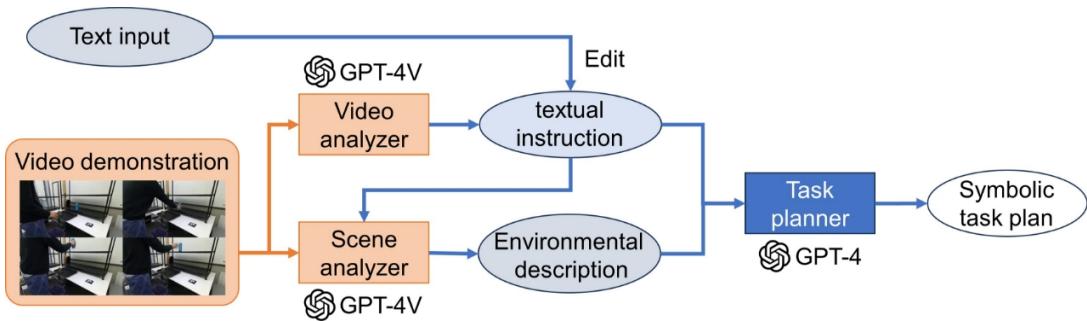
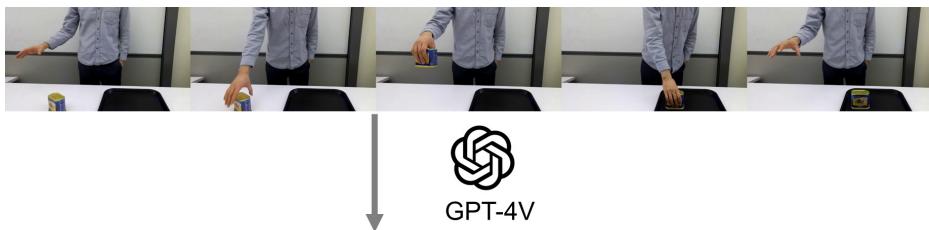
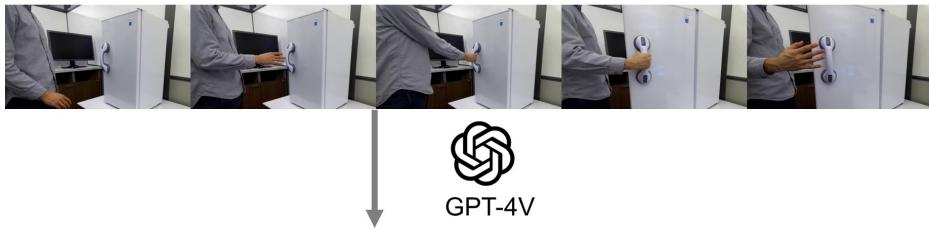


图 15：利用 GPT-4V 和 GPT-4 的多模式任务计划器概览。该系统处理视频演示和文本指令，生成供机器人执行的任务计划。



Reach for the can on the table, grab it, and then place it on the tray nearby.



Please move your hand to the fridge handle, grab it, and open the fridge door.

图 16：视频分析仪输出示例。以固定的时间间隔提取五帧图像并输入 GPT-4V。我们将在第 6.2.2 节介绍整个流程。

用户可以通过输入功能对 GPT-4V 的识别结果进行反馈，以便进行修正。这一功能旨在提高识别结果的准确性，同时也使操作更加稳健。

接下来，场景分析器根据指令和第一帧视频数据（或环境图像）将预期工作环境编译成文本信息。这些环境信息包括 GPT-4V 可识别的物体名称列表、物体的可抓取属性以及物体之间的空间关系。虽然这些计算过程在 GPT-4V 中是一个黑盒子，但这些信息是根据 GPT-4V 的知识和输入的图像/文本输出的。图 17 显示了场景分析器的输出示例。如图所示，GPT-4V 成功地选择了与操作相关的对象。例如，当人类在桌子上移动垃圾容器时，输出中会包含一张桌子，而在打开冰箱的任务中则忽略了桌子。这些结果表明，场景分析

器根据人类的操作对场景信息进行了编码。我们提示 GPT-4V 解释对象选择过程的结果以及这些选择背后的原因。在实践中，我们发现这种方法产生了合理的输出结果。最后，根据给定的文字说明和环境信息，任务规划器会输出一系列任务（Wake 等人，2023c）。

Instruction: "Reach for the can on the table, grab it, and then place it on the tray nearby."



```

"objects": ["<can>", "<table>", "<tray>"],
"object_properties": {
    "<can>": ["GRABBABLE"],
    "<table>": [],
    "<tray>": []
},
"spatial_relations": {
    "<can>": ["on(<table>)"],
    "<table>": [],
    "<tray>": ["on(<table>)"]
},
"your_explanation": "The human is expected to grab the can that is on the table and place it on the tray. The table and the tray are included as they are the surfaces involved in this action. The can is marked as GRABBABLE because it is the item being manipulated. Other objects and details in the environment not related to the action, such as the human, the background elements, and the clothing, are omitted because they do not influence the task of moving the can from table to tray."

```

Instruction: "Please move your hand to the fridge handle, grab it, and open the fridge door."



```

"objects": ["<fridge>", "<fridge_handle>"],
"object_properties": {
    "<fridge>": [],
    "<fridge_handle>": ["GRABBABLE"]
},
"spatial_relations": {
    "<fridge_handle>": ["on(<fridge>)"],
    "<fridge>": []
},
"your_explanation": "Human is moving their hand to the fridge handle, grabbing it, and opening the fridge door. Objects included are the fridge and fridge handle, as they are directly involved in the action. I omitted the computer monitor and other unmanipulated objects in the setting because they are not being interacted with for this task."

```

图 17：利用 GPT-4V 的场景分析器输出示例。我们将在第 6.2.2 节中介绍我们的整个流程。

机器人导航的化身代理。视觉语言导航（VLN）是指在真实的三维环境中，导航化身代理执行自然语言指令的任务。在三维环境中导航（Zhu 等人，2017a；Mirowski 等人，2016；Mousavian 等人，2018；Hemachandra 等人，2015）是在物理世界中发挥作用的移动智能系统的一项基本能力。在过去的几年中，人们提出了大量的任务和评估协议（Savva 等人，2017；Kolve 等人，2017；Song 等人，2017；Xia 等人，2018；Anderson 等人，2018a），总结如下（Anderson 等人，2018b）。VLN（Anderson 等人，2018a）侧重于真实三维环境中的语言导航。为了解决 VLN 任务，（Anderson 等人，2018a）建立了一个基于注意力的序列到序列基线模型。然后（Wang 等人，2018）引入了一种混合方法，结合了无模型和基于模型的强化学习（RL），以提高模型的泛化能力。最后，（Fried 等人，2018）提出了一种扬声器跟随者模型，该模型采用了数据增强、全景动作空间和针对 VLN 的修正波束搜索，在 Room-to-Room 数据集上确立了当前最先进的性能。在先前工作的基础上，我们在（Wang 等人，2019 年）中提出了针对 VLN 的强化跨模态匹配（RCM）。RCM 模型建立在（Fried 等人，2018 年）的基础上，但在许多重要方面有所不同：（1）RCM 将新颖的多奖赏 RL 与 VLN 的模仿学习相结合，而 Speaker-Follower 模型（Fried 等人，2018 年）仅使用监督学习，如（Anderson 等人，2018a）。（2）RCM 推理导航器执行跨模态接地，而不是单模态输入的时间注意机制。

(3) RCM 匹配批判器在架构设计上与 Speaker 类似，但前者用于为 RL 和 SIL 训练提供循环重构内在奖励，而后者则用于增强监督学习的训练数据。在 (Wang 等人, 2019) 一文中，我们研究了如何解决该任务的三个关键领导板问题：跨模态接地、不合理反馈和泛化问题。如图 18 所示，我们提出了一种新颖的 "强化跨模态匹配" (Reinforced Cross-Modal Matching) 方法，该方法可强化跨模态匹配。



图 18：VLN 任务的嵌入式代理演示（Wang 等人，2019 年）。图中显示了自上而下视图中的指令、本地视觉场景和全局轨迹。代理无法访问自上而下视图。路径 A 是遵循指令的演示路径。路径 B 和 C 是代理执行的两条不同路径。

通过强化学习（RL）实现局部和全局接地。特别是，我们使用匹配批评者提供内在奖励，以鼓励指令和轨迹之间的全局匹配，并使用推理导航器在本地视觉场景中执行跨模态接地。在 VLN 基准数据集上进行的评估表明，我们的 RCM 模型在 SPL 上比以前的方法高出 10%，达到了新的高水平。为了提高所学策略的通用性，我们进一步引入了自我监督模仿学习（SIL）方法，通过模仿自身过去的良好决策来探索未知环境。我们证明，SIL 可以逼近更好、更有效的策略，从而极大地缩小已见环境与未见环境之间的成功率差距（从 30.7% 降至 11.7%）。此外，在（Wang 等人，2019）一文中，我们引入了一种用于探索的自监督模仿学习方法，以明确解决泛化问题，而这一问题在之前的工作中并没有得到很好的研究。与这项工作同时进行的还有（Thomason 等人，2018；Ke 等人，2019；Ma 等人，2019a,b）从多方面研究了 VLN 任务，（Nguyen 等人，2018）引入了 VLN 任务的变体，即在需要时通过请求语言协助来寻找对象。需要注意的是，我们是第一个提出为 VLN 任务探索未知环境的人。

6.3 医疗保健

在医疗保健领域，LLM 和 VLM 可以充当诊断代理、病人护理助手，甚至是治疗辅助工具，但它们也承担着独特的责任。人工智能代理在改善患者护理和挽救生命方面潜力巨大，但同样危险的是，它们的滥用或仓促部署可能会危及全球成千上万或数百万人的生命。我们将讨论人工智能代理在医疗保健领域的一些前

景广阔的发展路线，同时也会讨论一些面临的主要挑战。

诊断代理。由于对医学专家的高需求以及 LLMs 帮助分流和诊断病人的潜力，将 LLMs 作为医疗聊天机器人用于病人诊断最近引起了极大关注（Lee 等人，2023 年）。对话代理，尤其是那些能够向来自不同患者群体的广大人群有效传达重要医疗信息的对话代理，有可能为历史上的弱势群体或边缘化群体提供公平的医疗服务。此外，世界各地的医生和医疗系统大多负担过重、资源不足，导致全球数亿人无法获得足够的医疗服务（世界卫生组织和世界银行，2015 年）。诊断制剂为改善数百万人的医疗保健提供了一个特别有利的途径，因为它们可以被制造成具有理解各种疾病的能力。

语言、文化和健康状况。初步结果表明，通过利用大规模网络数据，可以训练出具有医疗保健知识的 LMM (Li 等人, 2023f)。尽管这是一个令人兴奋的方向，但诊断代理的前景并非没有风险。我们将在下文中强调医疗环境中的幻觉风险，以及潜在的解决途径。

知识检索代理。在医疗领域，模型幻觉尤其危险，甚至可能导致严重的病人伤害或死亡，这取决于错误的严重程度。例如，如果病人错误地接受了一项诊断，而诊断结果表明他们并没有实际患有的疾病，这可能会导致灾难性的后果。这包括推迟治疗或治疗不当，或在某些情况下完全缺乏必要的医疗干预。未诊断或误诊病情的严重性会导致医疗费用增加，延长治疗时间，造成进一步的身体负担，在极端情况下还会造成严重伤害甚至死亡。因此，能够利用代理更可靠地检索知识 (Peng 等人, 2023 年) 或以检索为基础的方式生成文本 (Guu 等人, 2020 年) 的方法是大有可为的方向。将诊断代理与医学知识检索代理配对，有可能显著减少幻觉，同时提高诊断对话代理的响应质量和准确性。

远程医疗和远程监控。基于代理的人工智能在远程医疗和远程监控领域也有着巨大的潜力，它可以改善医疗服务的可及性，改善医疗服务提供者与患者之间的沟通，提高效率并降低医患频繁互动的成本 (Amjad 等人, 2023 年)。初级保健临床医生花费大量时间筛选患者信息、报告和电子邮件，而这些信息、报告和电子邮件往往与他们无关或没有必要查看。让支持代理帮助分流来自医生、患者和其他医疗服务提供者的信息，并帮助突出各方的重要信息，是大有可为的。通过让代理人工智能系统与患者、临床医生和其他人工智能代理进行协调，远程医疗保健和数字医疗保健行业有望迎来一场巨大的变革。

6.3.1 当前的医疗保健能力

图像理解。图 19 展示了 GPT-4V 等现代多模态代理在医疗保健领域的现有能力和局限性。我们可以看到，虽然 GPT-4V 对医院护理所涉及的设备和程序拥有丰富的内部知识，但它并不总是能对用户提出的更具说明性或诊断性的询问做出回应。

视频理解。我们从两个方面研究了 VLM 代理在医疗视频理解方面的性能。首先，我们研究了 VLM 代理在临床空间中识别重要病人护理活动的能力。其次，我们探讨了 VLM 在超声波等技术性较强的视频中的应用。具体来说，在图 20 中，我们展示了 GPT-4V 目前在医院护理和医疗视频分析方面的一些功能和局限性。

6.4 多模式代理

整合视觉和语言理解对于开发复杂的多模态人工智能代理至关重要。这包括图像字幕、视觉问题解答、视频语言生成和视频理解等任务。我们的目标是深入研究这些视觉语言任务，探索它们在人工智能代理方面的领先地位和机遇。

6.4.1 图像语言理解与生成

图像语言理解是一项涉及用语言解释给定图像中的视觉内容并生成相关语言描述的任务。这项任务对于开发能以更像人类的方式与世界互动的人工智能代理至关重要。其中最流行的任务包括图像标题（Lin 等人，2014 年；Sharma 等人，2018 年；Young 等人，2014 年；Krishna 等人，2016 年）、指代表达（Yu 等人，2016 年；Karpathy 等人，2014 年）和视觉问题解答（Antol 等人，2015 年；Ren 等人，2015 年；Singh 等人，2019 年）。

最近，人们提出了知识密集型视觉问题解答任务，如 OKVQA（Marino 等人，2019 年）、KB-VQA（Wang 等人，2015 年）、FVQA（Wang 等人，2017 年）和 WebQA（Chang 等人，2021 年）。多模态

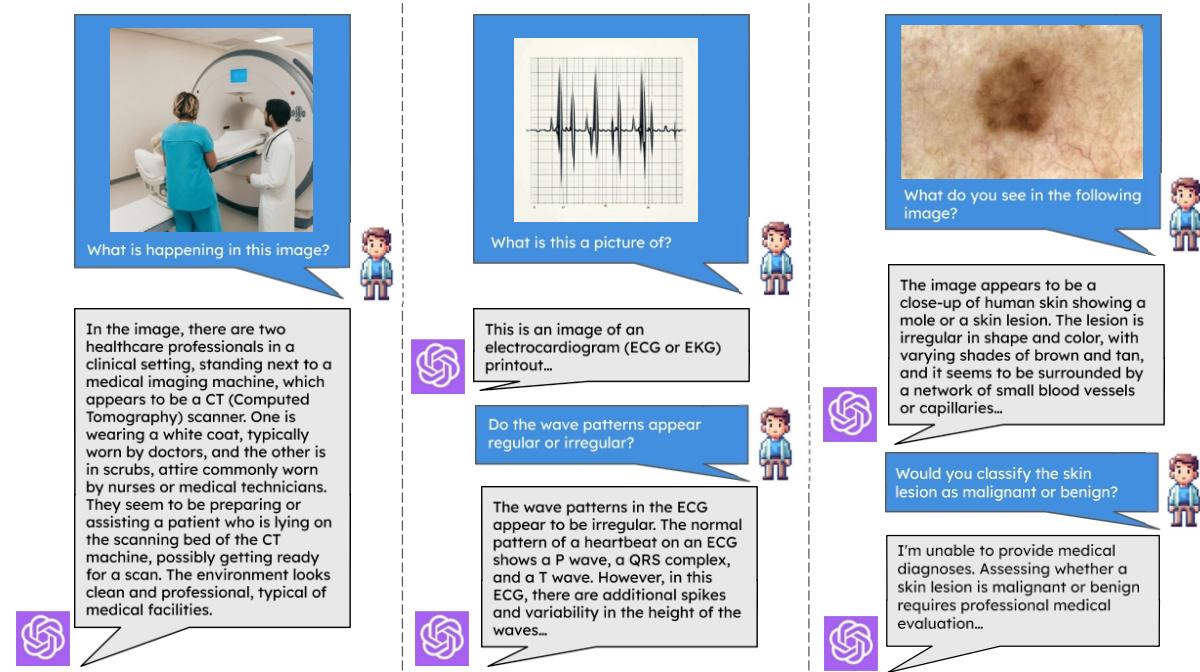


图 19：在医疗图像理解领域使用 GPT-4V 时的提示和回答示例。从左至右：（1）护士和医生进行 CT 扫描的图像；（2）不规则心电图扫描的合成图像；（3）来自 ISIC（Codella 等人，2018 年）皮肤病变数据集的图像。我们可以看到，GPT-4V 拥有丰富的医学知识，能够对医学图像进行推理。然而，由于安全训练的原因，它无法对某些医学图像做出诊断。

代理应该能够识别图像中的物体，理解它们之间的空间关系，生成关于场景的准确描述性句子，并利用推理技能处理知识密集型的视觉推理。这不仅需要物体识别能力，还需要对空间关系、视觉语义有深刻的理解，并能将这些视觉元素映射到语言结构中，与世界知识融为一体。

6.4.2 视频和语言理解与生成

视频语言生成。视频字幕或视频故事是为视频帧流生成一系列连贯句子的任务。受在视频和语言任务中成功使用递归大型基础模型的启发，代理驱动增强模型的变体在视频语言生成任务中显示出良好的效果。最根本的挑战在于，神经编码器-解码器模型的强大性能并不能很好地推广到视觉故事制作中，因为这项任务需要充分理解每幅图像的内容以及不同帧之间的关系。该领域的一个重要目标是创建一个代理感知的文本合成模型，该模型可以高效地对帧序列进行编码，并生成拓扑上连贯的多句子段落。

视频理解。视频理解将图像理解的范围扩展到动态视觉内容。这涉及到对视频中的帧序列进行解释和推理，通常还要结合附带的音频或文本信息。代理应能与视觉、文本和音频等各种模式进行交互，以展示其对视频内容的高级理解能力。该领域的任务包括视频字幕、视频问题解答和活动识别等。视频理解方面的领先优势是多方面的。它们包括视觉和语言内容的时间对齐、长帧序列的处理以及对随着时间推移而展开的复杂

活动的解释。在音频方面，代理可以处理口语、背景噪音、音乐和语音语调，从而理解视频内容的情绪、背景和微妙之处。



图 20：在医疗保健视频理解领域使用 GPT-4V 时的提示和响应示例。我们以 2x2 网格的形式输入示例视频，并叠加了标明帧顺序的文本。在前两个示例中，我们提示 GPT-4V 检查视频中的帧，以检测志愿者病人的临床床旁活动。在最后一个示例中，我们尝试提示 GPT-4V 评估一段超声心动图视频，但由于 GPT-4V 接受过安全培训，因此没有提供详细的回复。为了清晰起见，我们将描述相关活动的文字加粗，并缩写不必要的模型回答。为了保护个人隐私，我们将个人面孔涂成灰色。



图 21：交互式多模态代理包括四大支柱：交互、语音、视觉和语言。协同驾驶代理由不同的服务组成。1) 交互服务有助于为自动操作、认知和决策打造一个统一的平台。2) 音频服务将音频和语音处理整合到应用程序和服务中。3) 视觉服务可识别和分析图像、视频和数字墨水中的内容。4) 语言服务从结构化和非结构化文本中提取意义。

以前的工作主要是利用现有的视频语言训练数据建立视频基础模型（Li 等人，2020 年，2021b；Fu 等人，2022 年；Bain 等人，2021 年；Zellers 等人，2021 年，2022 年；Fu 等人，2023 年）。然而，由于这些数据集的有限性和不一致性，支持此类训练管道和功能十分困难。视频基础模型在设计时采用了遮蔽式和对比式预训练目标，之后再根据各自的任务进行调整。尽管这些模型在多模态基准测试中取得了不俗的成绩，但在纯视频任务（如动作识别）中却遇到了困难，因为它们依赖于从嘈杂的音频转录中建立的有限视频文本数据。这种局限性还导致大型语言模型缺乏一般所具备的鲁棒性和细粒度推理能力。

其他方法与图像语言理解中使用的方法类似，都是利用大型语言模型的强大推理能力和广泛知识来改进视频解释的不同方面。ChatGPT 和 GPT4 等纯语言模型或 GPT4-V 等图像语言模型简化了视频理解任务，它们将音频、视频和语言模式视为单独的可解释输入数据类型，并将代理定位为强大的开源模型。例如，（Huang 等人，2023c；Li 等人，2023g）通过使用开源视觉分类/检测/字幕模型将视频内容文本化，将视频理解转化为自然语言处理（NLP）的问题解答表述。（Lin 等人，2023）将 GPT4-V 与视觉、音频和语音方面的专业工具整合在一起，以促进复杂的视频理解任务，如在长视频中编写人物动作和行为脚本。

并行研究探索从大型模型中生成比例数据集，然后在生成的数据上应用视觉指令调整（刘等人，2023c；李等人，2023c；朱等人，2023）。随后，大量音频、语音和视觉专家感知模型被用于口头视频。语音由自动语音识别工具转录，视频描述和相关数据由各种标签、基础和字幕模型生成（Li 等人，2023g；Maaz 等人，2023；Chen 等人，2023；Wang 等人，2023f）。这些技术展示了在生成的数据集上对视频语言模型进行指导调整可如何提高视频推理和交流能力。

6.4.3 实验和结果

- **知识密集型模型：**如 INK（Park 等人，2022 年）和 KAT（Gui 等人，2022 年a）中介绍的那样，神经知识密集型任务包含人类注释的所需知识，以支持知识密集型检索任务。
- **多模态代理：**人们对变色龙（Chameleon）（Lu 等人，2023 年）和 MM-React（Yang 等人，2023 年c）等多模态语言模型的兴趣与日俱增。
- **视觉指令调整：** VCL（Gui 等人，2022b）、Mini-GPT4（Zhu 等人，2023）、MPLUG-OWL（Ye 等人，2023b）、LSKD（Park 等人，2023c）生成图像级指令调整数据集。

知识密集型代理。如图 22 和图 23 所示，基于知识的视觉问题解答和视觉语言检索任务是多模态机器学习中具有挑战性的任务，需要图像内容之外的其他知识。最近关于大规模转换器的研究主要集中在最大限度地提高模型参数存储信息的效率。本研究方向探索的是另一个方面：多模态变换器能否在决策过程中使用显性知识。基于变换器的预训练方法在跨模态隐式学习知识表征方面取得了显著的成功。然而，传统方法（主要是单模态方法）对知识检索和随后的答案预测进行了研究，提出了有关检索知识的质量和相关性以及使用内隐和外显知识整合推理过程的问题。为了解决这些问题，我们引入了知识增强转换器（KAT），该转换器在 2022 年 OK-VQA 开放域多模态任务中的表现比其他转换器高出 6%。KAT 采用编码器-解码器结构，将

来自 GPT3 的隐式知识与来自网站的显式知识相结合，并允许在生成答案时同时使用两种知识类型进行推理。此外，结合显性知识还能增强模型预测的可解释性。代码和预训练模型可在 <https://github.com/guilk/KAT> 上获取。

视觉语言转换器代理。接下来，我们将介绍 "从字幕训练视觉语言转换器" (VLC) 模型 (Gu 等人, 2022b)，这是一种完全使用图像字幕对进行预训练的转换器。尽管 VLC 只使用了一个简单的线性投影层进行图像嵌入，但与其他依赖对象检测器或有监督 CNN/ViT 网络的方法相比，VLC 在各种视觉语言任务中都取得了极具竞争力的结果。

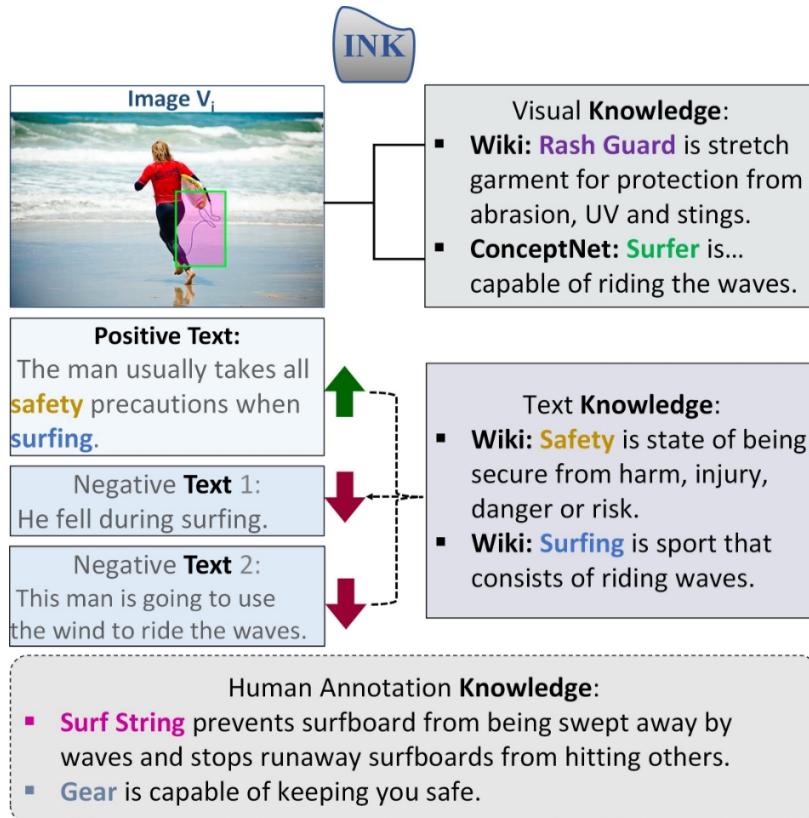


图 22：密集神经知识（INK）（Park 等人，2022 年）任务示例，该任务利用知识从一组候选文本中识别出与图像相关的文本。我们的任务涉及利用从网络和人类注释知识中检索到的视觉和文本知识。

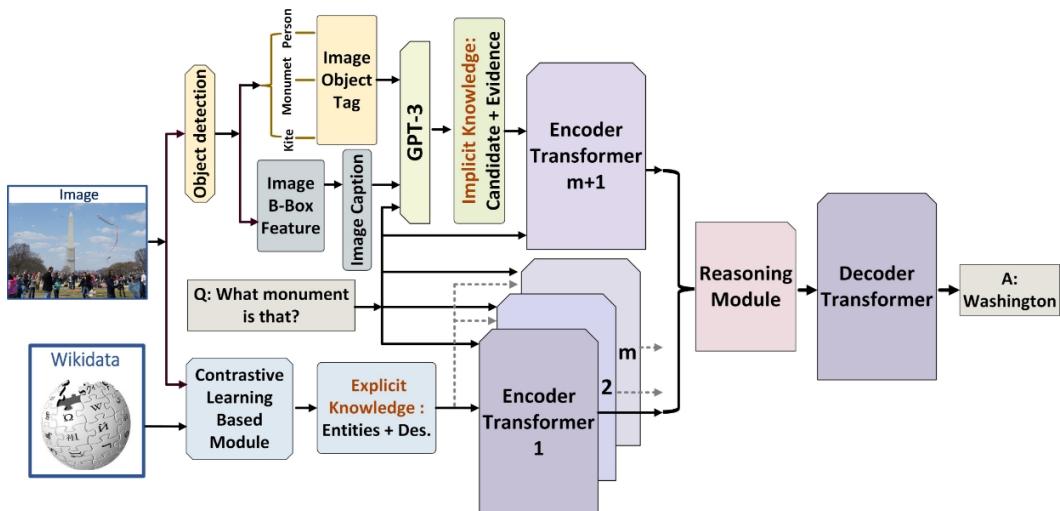


图 23：KAT 模型（Gui 等人，2022a）使用基于对比学习的模块从显性知识库中检索知识条目，并使用 GPT-3 检索具有支持证据的隐性知识。知识的整合由相应的编码转换器处理，并通过端到端训练与推理模块和解码转换器共同生成答案。

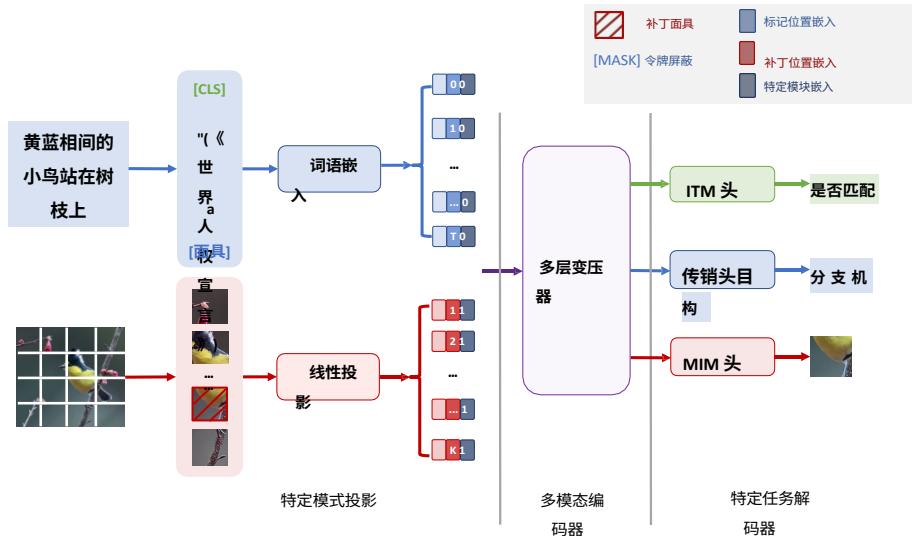


图 24：VLC 模型的整体结构（Gui 等人，2022b）。我们的模型由三个模块组成：(1) 特定模态投影。我们使用简单的线性投影来嵌入补丁图像，并使用单词嵌入层来嵌入标记化文本；(2) 多模态编码器。我们使用从 MAE（He 等人，2022 年）初始化的 12 层 ViT（Dosovitskiy 等人，2021 年）（ImageNet-1K，无标签）作为骨干；(3) 特定任务解码器。我们通过掩码图像/语言建模和图像-文本匹配来学习多模态表征，这仅在预训练中使用。我们使用双层 MLP 对下游任务的多模态编码器进行微调。重要的是，我们发现遮蔽图像建模目标在整个第二阶段预训练中都很重要，而不仅仅是视觉转换器的初始化。

通过广泛的分析，我们探索了 VLC 作为视觉语言转换器代理的潜力。例如，我们发现 VLC 的视觉表征对 ImageNet-1K 分类非常有效，而且我们的可视化结果证实，VLC 可以准确地将图像补丁与相应的文本标记相匹配。随着训练数据的增加，VLC 的性能也在不断提高，这凸显了开发大规模、弱监督、开放领域视觉语言模型的巨大潜力。

6.5 视频语言实验

为了了解将预先训练好的图像-LLMs 转换用于视频理解的实用性，我们对用于视频字幕的 InstructBLIP（Dai 等人，2023 年）进行了时间扩展和微调。具体来说，我们使用与“时间冻结”（Frozen in Time）（Bain 等人，2021 年）相同的分时空注意方案扩展 InstructBLIP 的视觉编码器（EVA-CLIP-G（Sun 等人，2023 年 b）），并在训练期间冻结 Q 形成器和 LLM（Flan-T5-XL（Chung 等人，2022 年））。我们冻结视觉编码器的所有空间层，同时在字幕训练期间保持时间层不冻结。这样，我们的模型就可以将图像和视频作为输入（与 InstructBLIP 的图像级性能相匹配）。我们在 WebVid10M（Bain 等人，2021 年）的 500 万个视频字幕子集上进行训练。我们在图 25 中展示了两个输出示例。然而，现有的代理无法完全理解视频内容中精确、精细的视觉细节。视觉指令调整方法也有类似的局限性，它们缺乏一般的、人类水平的感知能力，而这些能力仍有待多模态模型和代理来解决。

经过指令调整的模型在准确概括视频中的可见动作和识别动作（如图 25 中的“坐在长椅上的人”）方面表现

出色。但是，它们有时会添加错误的细节，如“人对着镜头微笑”，这暴露了在捕捉对话主题或视频氛围方面的不足，而这些元素对于人类观察者来说是显而易见的。这一不足凸显了另一个关键的局限性：忽略了音频和语音模式，而这些模式可以丰富视频理解的语境，有助于更准确的解读并防止此类错误表述。要弥合这一差距，就需要对现有模式进行全面整合，使多模式代理的理解能力达到与人类感知相近的水平，并确保以完全多模式的方法进行视频解读。

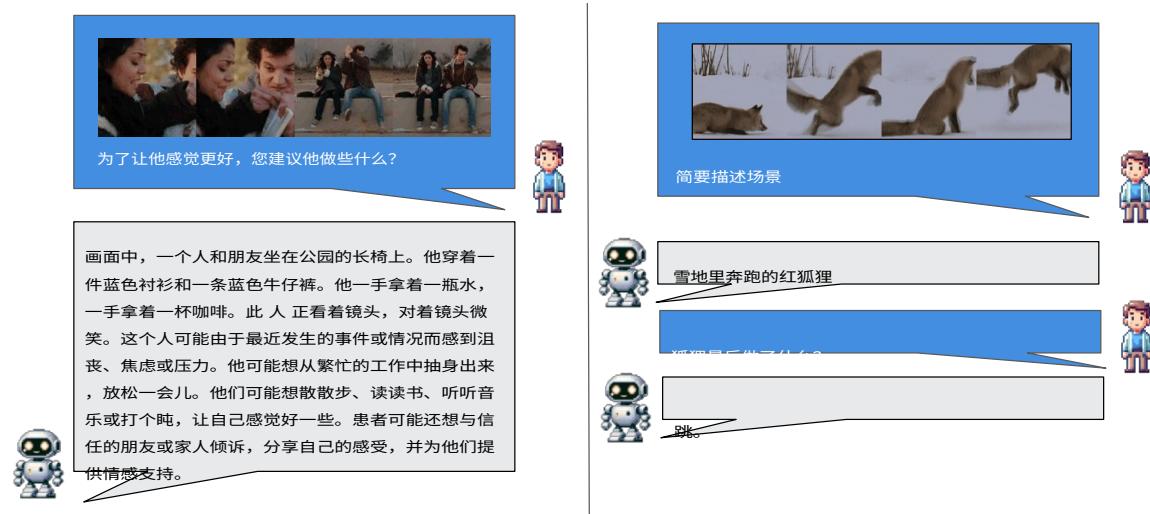


图 25：使用 InstructBLIP 视频微调变体时的提示和回答示例（方法见第 6.5 节）。我们的模型能够生成描述场景的长篇文字回复，并能回答与视频中事件的时间性相关的问题。

The diagram illustrates a multi-modal audio proxy, specifically a Video Chat Summary, GPT-4V Summary with Frame Captions, and GPT-4V Video Summary with Frame Captions + Audio.

Video Frames

GPT4-V Frame Captions

- A man with a serious expression stands holding a yellow measuring stick in a parking lot, with a car and trees in the background.
- A man in a "New York City" t-shirt holds a tall, yellow measuring stick upright in a sunny parking lot setting.
- A man is humorously biting a large yellow measuring stick while extending it with his left arm in a parking lot with a backdrop of greenery and a car.

VideoChat Summary

The video shows a man wearing a blue shirt who is holding a yellow pole and demonstrating how to use it. He is seen holding the pole and talking to the camera while explaining the process. The man is wearing a blue shirt and is holding a yellow pole. He is seen holding the pole and talking to the camera while explaining the process.

GPT-4V Summary with Frame Captions

In the video, a man appears in a series of sequences where he interacts with a large measuring stick in a parking lot. Initially, he stands with the stick at his side, presenting it like a novelty item. As the video progresses, he engages more actively with the stick, extending it outward as if sizing up some unseen object or distance. In a playful twist, he bites down on it while holding it horizontally, showcasing a humorous approach to what might be a mundane task. Throughout the video, the man's demeanor suggests he's involved in a light-hearted activity, possibly a humorous take on measurement and scale in an everyday suburban setting. The background with cars, greenery, and signage for a dental clinic remains consistent, anchoring the sequence of events in a single location and providing a casual, almost candid atmosphere to the activities unfolding.

GPT-4V Video Summary with Frame Captions + Audio

The video is a tutorial on how to measure the height of a tree using a simple method with a broomstick. The instructor demonstrates how to hold the broomstick perpendicular to the body and rotate it down so that when it's parallel to the eye, the length of the broomstick is adjusted accordingly. The stick is then flipped back up and held perpendicular again. The distance from the hand to the eye is said to be equal to the distance from the hand to the end of the stick. The instructor then explains how to use this measurement to create an isosceles triangle when sighting the bottom of the tree with one eye and aligning the top of the tree with the top of the stick. This alignment allows the measurer to determine that they are exactly as far from the tree as the tree is tall. Finally, the instructor advises to measure the distance from the point on the ground to the base of the tree with a tape measure to find the height of the tree, completing the measurement process.

图 26：第 6.5 节中描述的音频多模态代理。幻觉内容以红色标出。我们使用 GPT-4V 生成 1) 带有视频帧的视频聊天摘要；2) 带有帧标题的视频摘要；3) 带有帧标题和音频信息的视频摘要。

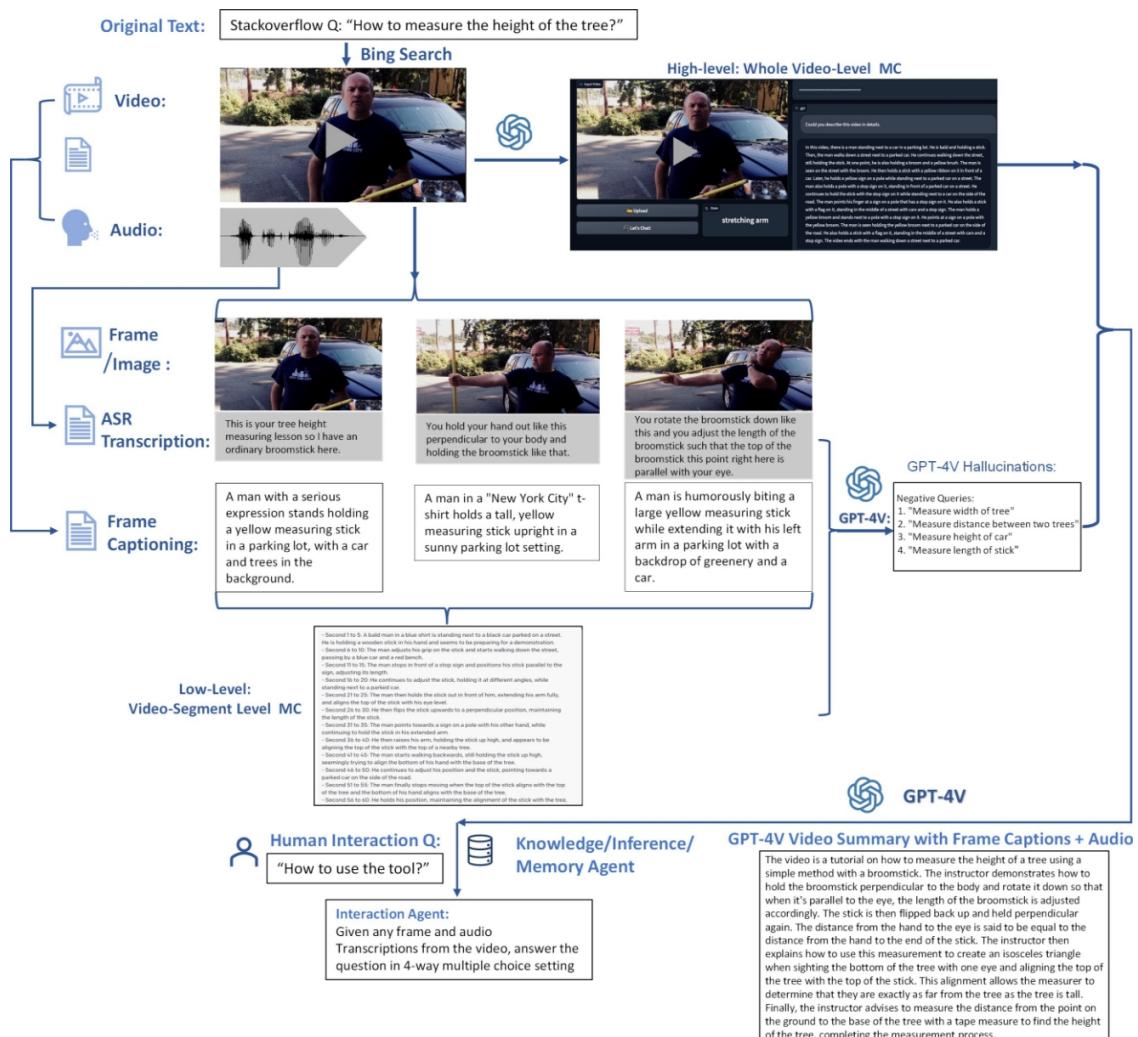


图 27：交互式多模态代理结合了视觉、音频和文本模式来理解视频。我们的管道可以挖掘出硬负面幻觉，从而为 VideoAnalytica 挑战赛提供高难度查询。关于交互式音视频语言代理数据集的更多相关细节，请参见第 9.2 节。

使用 GPT-4V 的音视频语言代理。然后，我们按照 (Lin 等人, 2023 年) 中概述的方法，评估了 GPT-4V 作为多模态代理的能力，该代理将视觉、音频和语音整合在一起，能够细致入微地精确理解视频。图 26 所示的结果比较了各种视频代理在视频摘要任务中的表现。视频指示调整模型 (Li 等人, 2023g) 提供了准确的内容，但在全面性和细节方面有所欠缺，缺少具体的动作，例如有条不紊地使用扫把测量树的高度。

为了提高视频描述的准确性，我们使用 GPT-4V 给帧添加字幕，而音频及其转录则来自 OpenAI Whisper 模型。然后，我们提示 GPT-4V 仅使用帧标题创建视频摘要，然后同时使用帧标题和音频转录创建视频摘要。最初，我们观察到，仅使用帧标题可能会导致捏造事件，例如第三段中有人咬住了一根棍子。在视频摘

要中，这些不准确的描述依然存在，例如 "他俏皮地一扭，咬住棍子，同时水平握住棍子"。在没有音频输入的情况下，代理无法纠正这些字幕错误，从而造成语义正确但视觉误导的描述。

然而，当我们向代理提供音频转录时，它能准确地描述内容，甚至能捕捉到详细的物理动作，如“握住与身体垂直的扫把并向下旋转”。这种程度的细节信息量明显更大，能让观众更清楚地了解视频的目的和关键细节。这些发现凸显了整合音频、视频和语言交互以开发高质量多模态代理的重要性。GPT-4V 为实现这种先进的多模态理解和交互奠定了良好的基础。

使用 GPT-4V 的嵌入式多模态代理。如图 27 所示，我们主要使用 StackOverflow 获取初始问题，然后使用“必应搜索”API 获取与问题相对应的相关视频和音频。接下来，我们主要使用 GPT-4V 获取相关文本信息和高级视频描述。另一方面，我们通过 ASR 将关键帧音频转为关键帧的低级片段描述。最后，我们使用 GPT-4V 生成令人信服的“幻觉”，作为视频问答任务的硬否定查询。我们支持视频当前帧中的互动和问题解答，以及整体高级视频描述的总结。在推理过程中，我们还通过网络搜索结合外部知识信息来提高回答能力。

GPT-4V 的主要提示信息如下。为清晰起见，整个提示信息以缩排方式显示，长达一页多。

GPT-4V 是一款辅助工具，可为能听不能看的视障人士提供描述性、信息性和全面的视频细节。GPT-4V 的任务是通过合成给定的注释来创建高质量、高密度的视频描述，并将其输出为 JSON 格式。具体来说，GPT-4V 将获得用于搜索视频的原始查询、视频标题、描述、音频转录以及视频中特定时间的潜在嘈杂描述。同一视频的不同片段被注释为 “[时间开始 - 时间结束（秒）] '文本'"。利用所有的转录和描述来推理视频中可能发生的确切细节和视觉演示。GPT-4V 将根据需要合并或分割时间戳，以提供最佳的视频分割效果。

对 GPT-4V 输出的期望：

1. 以动作为导向的描述：优先考虑音频中暗示的合理动作、动作和肢体演示，通过动态视觉线索丰富叙事内容。
2. 完整的视频覆盖：提供连续、一致的音频描述体验，覆盖视频的每一时刻，确保没有任何内容未被描述。
3. 简洁分段：用重点突出、简洁明了的片段（每个片段 1-2 句）进行描述，既能有效传达视觉动作，又不会过于详细。
4. 语境视听合成：将口语音频内容与推断出的视觉元素无缝融合，形成反映潜在屏幕活动的叙事。
5. 想象力丰富、似是而非的推测：在描述中加入有创意且可信的视觉细节，与音频相呼应，增强对场景的理解。
6. 精确的时间码对应：将描述性片段与相应的时间码对齐，确保推测的视觉细节与音频叙述的时间线同步。
7. 自信的叙述方式：在叙述时要有把握，就像所推测的视觉效果正在发生一样，给听众以信心。
8. 省略不合情理的细节：排除不符合所提供的视听信息所建立的背景的物体或事件的描述。

最终输出应采用 JSON 格式的结构，其中包含一个字典列表，每个字典详细记录一段视频。

最终输出应采用 JSON 格式的结构，其中包含一个字典列表，每个字典详细记录一段视频。

```
[ 'start': <start-time-in-seconds>, 'end': <end-time-in-seconds>, 'text': "<此处为您详  
细的单句视听描述>"]
```

MC 创建：我们的任务是为视频到文本检索任务创建多选题，通过查看标题和阅读音频转录即可轻松解决这些问题。为此，我们将给出原始查询，以获取视频、描述、音频转录以及视频中特定时间段可能存在的嘈杂描述。

- 音频转录格式: -[开始-结束时间 (秒)] "转录"
- 噪音描述格式: -[时间 (秒)] "描述"

我们恳请 GPT-4V 生成四个查询，其中主查询与视频内容一致，其他三个否定查询与我们的主查询有细微差别。在选择主查询时，不应简单地听取音频转录，例如，原始查询文本包含在音频转录中。底片应与视频内容密切相关，但又不完全一致，需要对视频有直观的理解才能区分。例如，以细微的方式修改语义，使人们在选择原始查询时需要观看视频而不仅仅是收听。用类似标题的语句编译四个查询，第一个是重新表述的原始查询。

逐步思考如何利用视频中的信息提出否定性陈述。并证明负面查询是不正确的，但仍然是引人注目的选择，需要对视频有细致入微的理解。以及人类如何不会意外地选择否定而不是原始查询。

最后，我们以以下分析和 4 个查询的格式介绍工作。无需生成如何翻译原始查询。

- 视频分析: XXX
- 查询: [查询 1、查询 2、查询 3、查询 4]
- 理由: xxx

6.6 NLP 代理

6.6.1 法学硕士代理

几十年来，识别任务指令并采取行动一直是交互式人工智能和自然语言处理领域的基本挑战。随着近年来深度学习的发展，人们对联合研究这些领域以改善人机协作的兴趣与日俱增。除其他外，我们确定了三个具体方向，以改进基于语言的代理：

- **工具使用和知识库查询。**这一方向强调了将外部知识库、网络搜索或其他有用工具整合到人工智能代理推理过程中的重要性。通过利用各种来源的结构化和非结构化数据，人工智能代理可以增强其理解能力，并提供更准确、更能感知上下文的响应。此外，它还能提高人工智能代理在面对陌生场景或查询时主动寻找信息的能力，确保做出更全面、更明智的回应。这方面的例子包括 Toolformer (Schick 等人, 2023 年) 和 Retrieve What You Need (Wang 等人, 2023 年g)。
- **改进代理推理和规划。**增强代理的推理和规划能力对于有效的人机协作至关重要。这就需要开发能够理解复杂指令、推断用户意图和预测潜在未来情景的模型。要做到这一点，可以像 ReAct (姚等人, 2023a) 那样要求代理反思过去的行动和失败，或者将代理的思维过程结构化为一种搜索形式 (姚等人, 2023b)。通过模拟不同的结果并评估各种行动的影响，代理可以做出更明智的情境感知决策。
- **结合系统和人类反馈。**人工智能代理经常在两种主要环境中工作：一种是提供明确行动效果信号的

环境（系统反馈），另一种是与人类合作的环境，人类可以提供口头批评（人类反馈）。这一方向强调了对自适应学习机制的需求，该机制允许代理改进其策略并纠正错误，例如在 AutoGen 中（Wu 等人，2023 年）。从不同的反馈来源不断学习和适应的能力，可确保代理始终对用户需求有帮助并与之保持一致。

6.6.2 普通法律硕士代理

几十年来，识别和理解代理内容和自然语言一直是交互式人工智能和自然语言处理领域的一项基本挑战。随着深度学习技术的不断进步，人们越来越关注

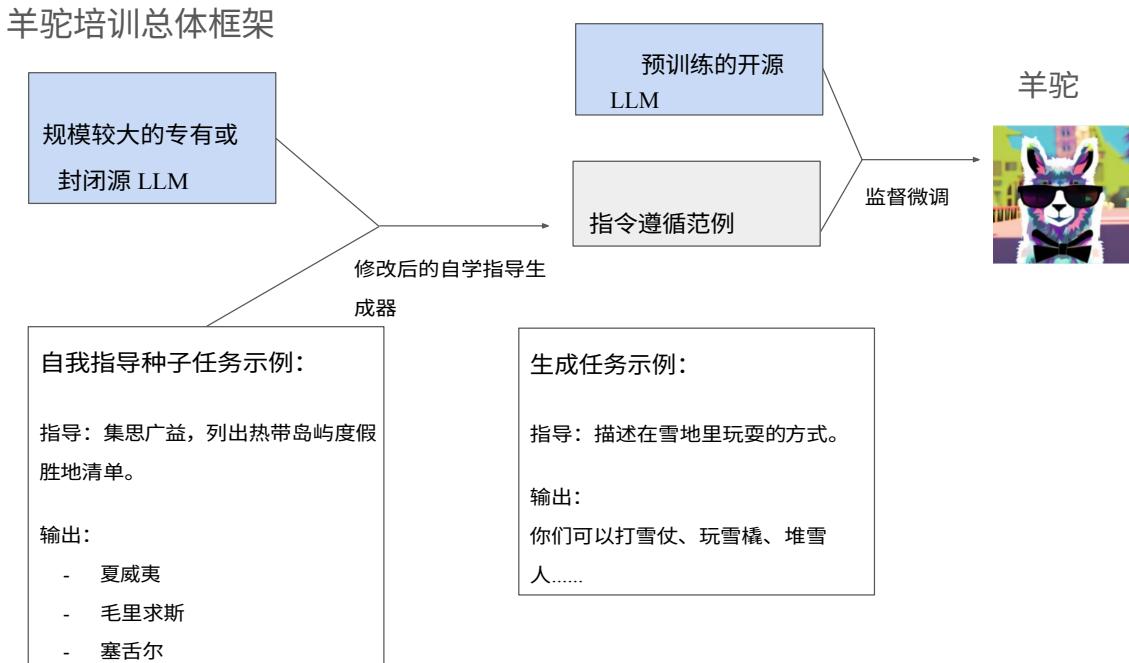


图 28：用于训练 Alpaca 模型的训练配方（Taori 等人，2023 年）。在高层次上，现有的 LLM 被用来从较小的种子任务集合中生成大量的指令跟随示例。然后，生成的指令跟踪示例将被用于对基础模型权重可用的 LLM 进行指令调整。

为了深入了解代理规划或人类对知识推理和自然语言生成的反馈，我们正在联合研究这两个领域。这些是许多无人机交互代理的关键组成部分，如“自动生成”（Wu 等人，2023 年）和“检索你所需要的”（Wang 等人，2023 年^g）。

6.6.3 遵循指导的 LLM 代理人

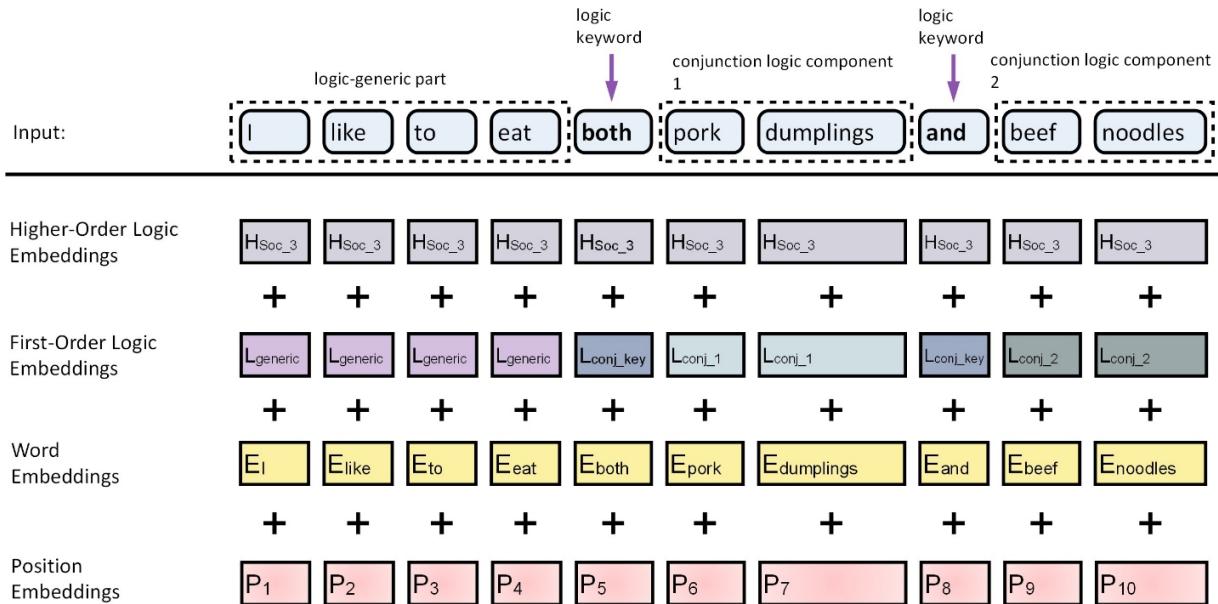
此外，创建可接受训练以有效遵从人类指令的 LLM 代理已成为一个重要的研究领域。最初的模型使用人类反馈来训练代理奖励模型，以模拟人类的偏好，这一过程被称为“人类反馈强化学习”（RLHF）（欧阳等人，2022 年）。这一过程产生了 InstructGPT 和 ChatGPT 等模型。为了更有效地训练遵循指令的 LLM 代理而不需要人类标签，研究人员开发了一种更有效的指令调整方法，直接根据指令/响应对训练 LLM 代理。⁶或 Alpaca 等 LLM（Taori et al.）我们在图 28 中展示了整个 Alpaca 训练流水线。

6.6.4 实验和结果

尽管会话系统和自我反馈系统的采用率越来越高，但这些形式的人工智能在根据自身的隐含知识生成符合事实的正确回应方面仍然表现不佳，因此在推理时往往需要使用网络搜索和知识检索机制等外部工具来增强其回应。解决这一问题将有助于在许多现实生活应用中为用户创造更具吸引力的体验。在社交会话（如

Instagram 和 Facebook 等社交媒体平台上的会话) 或问答网站 (如 Ask 或 Quora) 中, 人们通常通过一系列评论和网络搜索与讨论相关的信息和知识来与他人互动。因此, 在这种情况下生成对话转向的任务不是简单地利用传统的 NLP 模型和任务, 而是利用代理通过反映知识搜索和获取的智能行为来生成对话 (Peng 等人, 2023 年)。这样, 智能代理

《多莉 2.0》博文链接



These logic embeddings vectors are fully parameterized and are fully trainable during the fine-tuning training process on dialogue summarization datasets. The only restrictions that we put on their values are:

- Logic embeddings with same color (i.e. same logic role and type) must have the same vector values;
- Logic embeddings with different colors (i.e. different logic roles and types) can have different vector values.

图 29：逻辑转换器代理模型（Wang 等人，2023e）。我们在基于转换器的抽象摘要模型中集成了逻辑推理模块，以赋予逻辑代理对文本和对话逻辑进行推理的能力，从而使其能够生成更高质量的抽象摘要并减少事实性错误。

NLP 任务通过在对话过程中添加明确的知识搜索和检索步骤，扩展了任务描述并提高了响应的可解释性。将这些网络搜索和检索代理作为对话期间的反馈，将有助于进一步加深人类与代理之间的社会互动（Wang 等人，2023e）。如图 29 所示，我们为转换器语言模型引入了一种新的建模范式，即从输入文本中检测和提取重要的逻辑结构和信息，然后通过精心设计的多层分级逻辑投影将其整合到输入嵌入中，从而将逻辑结构注入预训练的语言模型，作为一种 NLP 代理。（Wang 等人，2023e）提出了一种新方法，通过逻辑检测、逻辑映射和分层逻辑投影相结合的方式，为转换语言模型构建逻辑感知输入嵌入，然后开发出一种相应的新建模范式，可将所有现有转换语言模型升级为逻辑转换器，从而持续提升其性能。通过更深入地理解文本的逻辑结构，所提出的逻辑转换器代理可持续实现优于其基准转换器模型的性能。对于人类用户来说，通过基于代理的对话和信息检索之间的协调，提供有意义和有趣的对话，这些方面往往更为重要。本专题将深入探讨自然语言处理，讨论在使 LLM 更具代理功能、更适合各种以语言为中心的任务方面取得的进展和领先优势。

开放域问题解答（QA）系统通常采用“检索-阅读”（retrieve-then-read）模式，即使用检索器从大型语料库中检索相关段落，然后由阅读器根据检索到的段落和原始问题生成答案。在（Wang 等人，2023g）一文中，我们提出了一个简单而新颖的相互学习框架，通过一个名为知识选择器代理的中间模块来提高“检索-阅读”

"元模型的性能。细粒度知识选择器融入了 "检索-阅读" 范式，其目标是构建一小部分保留问题相关信息的段落子集。如图 30 所示，知识选择器代理是作为我们新颖的相互学习框架的一个组成部分进行训练的，该框架对知识选择器和阅读器进行迭代训练。我们采用了一种简单而新颖的方法，即利用策略梯度来优化知识选择器代理，利用读者的反馈来训练它选择一个小而精的知识子集。

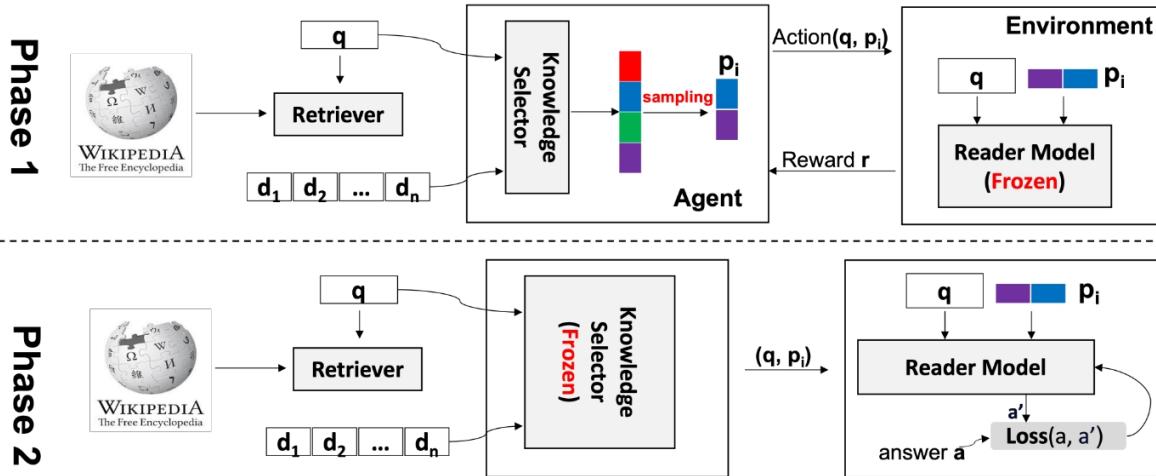


图 30：一个拟议的 NLP 代理 (Wang 等人, 2023g) 相互学习框架的架构。在每个纪元中，第 1 阶段和第 2 阶段交替执行。在第 1 阶段，阅读器模型的参数保持不变，只更新知识选择器的权重。相反，在第 2 阶段，阅读器模型的参数进行调整，而知识选择器的权重保持不变。

信息丰富的段落集。这种方法避免了野蛮搜索或人工设计的启发式方法，而且不需要任何有注释的查询-文档对作为监督。我们的研究表明，对阅读器和知识选择器代理进行迭代训练，可以在一些公开领域的问题解答基准测试中获得更好的预测性能。

7 跨越模式、领域和现实的代理人工智能

7.1 促进跨模态理解的代理

由于缺乏包含视觉、语言和代理行为的大规模数据集，多模态理解是创建通用人工智能代理的一大挑战。更广泛地说，人工智能代理的训练数据通常是特定模式的。这导致大多数现代多模态系统使用冻结子模态的组合。一些著名的例子有 Flamingo (Alayrac 等人, 2022 年)、BLIP-2 (Li 等人, 2023 年 c) 和 LLaVA (Liu 等人, 2023 年 c)，它们都使用了冻结 LLM 和冻结视觉编码器。这些子模块分别在不同的数据集上进行训练，然后训练适应层，将视觉编码器编码到 LLM 嵌入空间中。为了进一步提高人工智能代理的跨模态理解能力，很可能需要改变使用冻结 LLM 和视觉编码器的策略。事实上，RT-2 是最近推出的一种视觉语言模型，它能够在机器人领域内采取行动，当针对机器人和视觉语言任务联合调整视觉编码器和 LLM 时，其性能得到了显著提高 (Brohan 等人, 2023 年)。

7.2 跨域理解代理

创建通用代理所面临的一个关键挑战是，不同领域的视觉外观和行动空间各不相同。人类一旦熟悉了现实世界、视频游戏以及机器人和医疗保健等专业领域的具体细节，就有能力解读来自这些领域的图像和视频

。然而，现有的 LLM 和 VLM 在训练数据和应用的不同领域之间往往存在显著差异。值得注意的是，当试图开发一种能有效学习跨领域多个控制系统的单一策略时，训练代理模型来预测特定行动是一个相当大的挑战。一般来说，在特定领域内应用系统时，大多数现代作品所采用的方法是从预训练的基础模型开始，然后针对每个特定领域对单独的模型进行微调。这种方法无法捕捉到各领域之间的共性，导致用于训练的数据总量减少，而无法充分利用各领域的数据。

7.3 跨模态和跨现实的交互式代理

开发能够在不同现实环境中成功理解和执行任务的人工智能代理是一项持续的挑战，最近在图像和场景生成方面取得了一些成功（Huang 等人，2023a）。特别是，由于视觉上的差异和独立的环境物理特性，人工智能代理要同时理解现实世界和虚拟现实环境具有挑战性。在跨现实的背景下，当使用模拟训练的策略处理现实世界的数据时，模拟到现实的转换是一个特别重要的问题，我们将在下一节讨论这个问题。

7.4 模拟到真实的传输

使在模拟环境中训练的模型能够在现实世界中部署的技术。嵌入式代理，尤其是基于 RL 策略的代理，通常是在模拟环境中训练出来的。这些模拟环境并不能完全复制真实世界的特征（如干扰、光线、重力和其他物理特性）。由于模拟与现实之间存在这种差异，在模拟环境中训练出来的模型在应用于现实世界时往往难以取得良好的效果。这个问题被称为 "模拟到现实" 问题。要解决这个问题，可以采取以下几种方法：

- **领域随机化：**领域随机化是一种在模拟环境中随机改变参数（如物体外观、传感器噪声和光学特性）来训练模型的技术，以预测真实世界的不确定性和变化（Tobin 等人，2017 年）。例如，在训练基于 RL 的抓取技能时，在物体形状中引入随机性可导致一种能够适应不同形状物体的策略（Saito 等人，2022 年）。
- **领域适应：**领域适应或领域转移是一种技术，它通过使用大量模拟图像和较小的真实世界图像集来训练模型，从而弥合模拟领域和真实世界领域之间的差距。在实际环境中，由于难以准备跨领域的配对图像，因此采用了无配对图像到图像的转换方法，如 Cy-cleGAN（Zhu 等人，2017b）。还有一些强化学习的增强版本，包括 RL-CycleGAN（Rao 等人，2020 年），以及用于模仿学习的 RetinaGAN（Ho 等人，2021 年）。
- **改进模拟：**逼真的模拟是模拟到现实转换的关键。这项工作的一部分是通过系统识别技术实现的（Zhu 等人，2017c；Allevato 等人，2020），该技术旨在识别模拟参数，以模仿真实世界的环境。此外，使用逼真的模拟器对基于图像的强化学习也很有效（Martinez-Gonzalez 等人，2020；Müller 等人，2018；Shah 等人，2018；Sasabuchi 等人，2023）。

随着各种方法的不断发展，模拟到现实的转换仍然是 "化身代理" 研究的核心挑战。理论和实证研究对于进一步推动这些技术的发展至关重要。

8 人工智能代理的持续和自我完善

目前，基于基础模型的人工智能代理有能力从多个不同的数据源进行学习，这使得训练数据的来源更加灵活。这样做的两个主要结果是：(1) 用户和基于人类的交互数据可用于进一步完善和改进人工智能代理；(2) 现有的基础模型和模型工件可用于生成训练数据。我们将在下面的章节中详细讨论上述两点，但我们要指

出的是，由于目前的人工智能代理在很大程度上与现有的预训练基础模型相绑定，因此它们通常无法从与环境的持续交互中学习。我们认为这是一个令人兴奋的未来方向，Bousmalis 等人的初步研究表明，用于机器人控制的自我改进型代理能够在没有监督的情况下通过环境互动不断学习和改进（[Bousmalis 等人，2023 年](#)）。

8.1 人机交互数据

使用人机交互数据的核心理念是利用大量的代理与人机交互数据来训练和改进代理的未来迭代。有几种策略可用于从人机交互中改进代理。

- **额外的训练数据** 也许人与代理交互最简单的用法就是将交互示例本身作为未来代理迭代的训练数据。
◦ 这通常需要过滤

将成功的代理示例与不成功的交互示例区分开来的策略。筛选可以是基于规则的（例如，达到某种预期的最终目标状态）、基于模型的（例如，对成功与不成功的交互进行分类），也可以是在对交互示例进行事后检查和/或修改后手动选择的。

- **人类偏好学习** 在与用户交互过程中，代理系统可以向用户提示几种不同的模型输出结果，并允许用户选择最佳输出结果。ChatGPT 和 GPT-4 等 LLM 通常采用这种方法，用户可以从几种输出中选择一种最符合其偏好的输出。
- **安全培训（红队）** 在人工智能代理的背景下，红队指的是有一个专门的对手团队（人类或计算机），试图利用和暴露人工智能代理系统中的弱点和漏洞。虽然具有对抗性质，但“红队”通常被用作了解如何改进人工智能安全措施和减少有害输出的一种手段。其核心原则是发现诱导不需要的代理输出的一致方法，以便在明确纠正这种行为的数据上对模型进行训练。

8.2 基础模型生成的数据

随着学术界和工业界制作的功能强大的基础模型作品的出现，人们开发了各种方法，利用各种提示和数据配对技术从这些作品中提取和生成有意义的训练数据。

- **LLM 指令调谐** 从 LLM 生成指令跟随训练数据的方法，可以在大型专有 LLM 输出的基础上对小型开源模型进行微调（Wang 等人，2022b）。例如，Alpaca（Taori 等人，2023 年）和 Vicuna（Zheng 等人，2023 年）是基于开源 LLaMA 系列（Touvron 等人，2023 年）的 LLM，已根据 ChatGPT 和人类参与者的各种输出进行了调整。这种教学调整方法可被视为一种知识提炼形式，其中较大的 LLM 可作为较小的学生模型的教师模型。重要的是，尽管 LLM 的教学调整已被证明能将教师模型的写作风格和一些指令遵循能力转移到学生模型上，但教师模型和学生模型的事实性和能力之间仍存在巨大差距（Gudibande et al.）
- **视觉语言对** 最近有许多研究试图通过为视觉内容自动生成标题和其他文本，来增加视觉语言模型可用的预训练数据的多样性。例如，LLaVA（Liu 等人，2023c）使用了 150,000 个主要由 LLM 生成的文本和视觉输入的指令跟随行为示例。其他研究表明，使用 VLM 为图像重新标注字幕可以提高图像生成模型的训练数据和后续质量（Segalis 等人，2023 年）。在视频理解领域，使用 VLMs 和 LLMs 对视频进行重捕捉已被证明可提高在重捕捉视频上训练的后续 VLMs 的性能和质量（Wang 等人，2023f；Zhao 等人，2022）。

9 代理数据集和排行榜

为了加速这一领域的研究，我们分别针对多代理游戏和代理视觉语言任务提出了两个基准。我们将发布两个新数据集--“CuisineWorld”和“VideoAnalytica”--以及一组基准模型，鼓励参与者探索新模型、新系统，并在我们的排行榜测试集上提交他们的结果。

9.1 用于多代理游戏的 "CuisineWorld" 数据集

CuisineWorld 是一款基于文本的游戏，让人想起《Overcooked》！它为人工智能驱动的代理提供了一个合作和协同游戏的平台。该数据集将测试多代理系统的协作效率，深入了解 LLM 和其他系统在动态场景中的协作能力。特别是，该数据集将重点关注代理对目标的理解程度，以及代理之间的协调能力。本数据集支持两种模式：集中调度模式和分散模式。参与者可以选择一种游戏模式，并向我们的排行榜提交作品。

9.1.1 基准

在我们的竞赛中，我们将发布一个基准，即 CuisineWorld 基准，其中包括一个包含可扩展任务定义文件的文本界面，以及一个用于多代理交互和人机交互的界面。我们介绍了游戏交互任务，其目标是生成相关的、适当的、能最大限度提高协作效率的多代理协作策略。我们用提出的评价指标来评估协作效率：CoS。

CuisineWorld "数据集由微软、加州大学洛杉矶分校和斯坦福大学收集。竞赛的目的是探索不同的、现有的和新颖的基础 LLM 和交互技术在这一基准中的表现，并为多代理游戏基础设施任务建立强大的基准。

CuisineWorld 的数据集包括

- 选择定义明确的多代理协作任务。
- 促进代理互动的应用程序接口系统。
- 自动评估系统。

(下载数据集的链接很快就会公布，本文也将在此更新)。

9.1.2 任务

- 我们提供了一个名为 Microsoft MindAgent 的数据集和相关基准，并相应地向研究界发布了一个名为 "CuisineWorld "的数据集。
- 我们将提供基准，对提交的 "MindAgent "算法进行评估和排序。我们还将提供使用流行基础设施生成的基准结果。

9.1.3 衡量标准和评判

多代理协作效率的质量由新的 "cos "自动度量（来自 MindAgent ([Gong 等人，2023a](#))）决定。该指标的最终评分是根据多代理系统在所有任务中的协作效率指标的平均值计算得出的。人类评估者将被要求对单个响应进行评分，并对用户与代理互动的参与度、广度和整体质量进行主观判断。

9.1.4 评估

- **自动评估。**我们计划发布一个排行榜，从发布日期（待定）开始，注册参与者将被要求提交与数据集 "CuisineWorld"（我们公开发布的排行榜数据集）相关的任务结果。结果提交将于结束日期（待定）截止。每个团队都必须提交他们在测试集上生成的结果，以便对 "cos "指标进行自动评估。
- **在我们的排行榜上进行人工评估。**排行榜参与者需要提供由本地评估脚本生成的提交文件。我们将使用 evalAI 系统检查提交文件，并选择性地重新运行顶级挑战者的代码。因此，参赛团队在提交代码的同时还必须提交一份自述文件，说明如何运行其代码。人工评估将由组织团队执行。
- **获奖者公布。**我们将公布获奖者名单，并在排行榜上公布参赛作品的最终评分。

9.2 音视频语言预培训数据集。

我们介绍 VideoAnalytica：分析性视频演示理解的新基准。VideoAnalytica 专注于利用视频演示作为辅助工具，更好地理解长篇教学视频中蕴含的复杂、高级推理。其目的是评估视频语言模型的认知推理能力，使其超越单纯的识别任务和基本理解，进而对视频进行更复杂、更细致的理解。最重要的是，VideoAnalytica 强调多种模式的整合，例如

VideoAnalytica 包含音频、视频和语言，以及模型应用特定领域知识的能力，以便对视频中呈现的信息进行上下文关联和解释。具体来说，VideoAnalytica 包括两项主要任务：

1. **视频文本检索：**这项任务涉及从教学视频中准确检索相关文本。其难点在于如何区分相关信息和无关信息，因此需要深入理解视频内容，并对演示进行分析，以检索出正确的查询。为了进一步增加这些任务的复杂性，我们在数据集中引入了由大型语言模型生成的硬否定词。我们对生成的否定词进行人工验证，并删除使任务无效和不公平的实例（例如否定词有效）。
2. **视频辅助信息问题解答：**这项任务要求模型根据从视频中提取的信息回答问题。重点是需要分析推理和全面理解视频演示的复杂问题。

为了促进用于分析视频理解的音频视频语言代理的开发，我们从 VideoAnalytica 中引入了两个任务的基准排行榜。

- 排行榜参与者需要提交他们的解决方案以供评估。评估将基于模型在两项任务中的表现，结果将显示在排行榜上。参赛者需提交其代码，并详细说明其方法和手段。
- 伦理考虑：排行榜侧重于理解和解释视频内容，这有可能被用于监控或其他侵犯隐私的应用。因此，考虑技术的道德影响和潜在滥用至关重要。我们鼓励参赛者在提交作品时考虑这些方面，并促进人工智能的道德应用。

10 更广泛的影响声明

本文及相关论坛⁷旨在成为创新研究的催化剂，促进合作，推动下一波人工智能应用的发展。通过关注多模态代理，我们强调了人与人工智能互动、排行榜和解决方案的未来方向。我们详细介绍了我们为更广泛的社区做出重大贡献的三种方式。

首先，我们希望我们的论坛能让人工智能研究人员从游戏、机器人、医疗保健和长视频理解等领域的实际问题出发，开发出解决方案。具体来说，在游戏领域开发多模态代理可以带来更加身临其境和个性化的游戏体验，从而改变游戏产业。在机器人领域，自适应机器人系统的开发可彻底改变从制造业到农业的各个行业，从而有可能解决劳动力短缺问题并提高效率。在医疗保健领域，使用 LLM 和 VLM 作为诊断代理或病人护理助手，可以提高诊断的准确性，改善病人护理，增加医疗服务的可及性，特别是在服务不足的地区。此外，这些模型解释长视频的能力可能会产生深远的应用，从加强在线学习到改善技术支持服务。总之，我们论坛所涉及的主题将对全球各行各业和人类产生重大的下游影响。

其次，我们希望我们的论坛能成为人工智能从业人员和研究人员的宝贵资源，成为一个探索和深入理解在各种环境和情况下实施人工智能代理所带来的多样而复杂的排行榜的平台。例如，这种探索包括了解在为医疗诊断等专业领域开发人工智能代理系统时与之相关的具体限制和潜在危险。在这一领域，人工智能行为中的危险幻觉等问题可能会带来重大风险，这就凸显了精心设计和测试的重要性。然而，在考虑为游戏行业

精心设计的人工智能代理时，这些特定的排行榜可能并不同样相关或引人注目。在此类娱乐领域，开发人员可能会优先考虑解决不同的障碍，例如人工智能需要进行更开放的生成并表现出创造力，动态适应不可预测的游戏场景和玩家互动。通过参加论坛，与会者将深入了解这些不同的环境如何决定人工智能开发的重点和方向，以及如何最好地定制人工智能解决方案以满足这些不同的需求并克服相关的排行榜。

⁷<https://multimodalagentai.github.io>

第三，我们活动的各种元素，包括专家演讲、内容丰富的海报，特别是我们两个排行榜的获奖者，都将对多模态代理领域的最新和重要趋势、研究方向和创新理念提供一个实质性而又简洁的概述。这些演讲将囊括重要的研究成果和发展动态，揭示多模态代理人工智能领域的新系统、新理念和新技术。这些知识不仅有利于我们论坛的与会者加深对这一领域的理解和专业知识，而且也是一个充满活力、内容丰富的资源库。访问我们论坛网站的人可以利用这些信息库，发现和了解引领多模态代理人工智能未来的前沿进展和创意。我们致力于为该领域的新手和老手提供一个有用的知识库。通过利用这些资源，我们希望与会者和在线访问者都能随时了解正在塑造令人兴奋的多模态代理人工智能景观的变革和新方法。

11 伦理方面的考虑

多模态代理人工智能系统有许多应用。除交互式人工智能外，基础多模态模型还有助于推动机器人和人工智能代理的内容生成，并协助生产力应用，帮助重新播放、转述、行动预测或合成三维或二维场景。人工智能代理的根本性进步有助于实现这些目标，而对如何在模拟现实或真实世界中建立具身和移情模型的更深入了解将使许多人受益。可以说，这些应用中有许多都能带来积极的益处。

不过，这项技术也可能被坏人利用。生成内容的人工智能代理系统可以用来操纵或欺骗人们。因此，按照负责任的人工智能准则开发这种技术非常重要。例如，明确告知用户内容是由人工智能系统生成的，并为用户提供控制功能，以便定制这样的系统。代理人工智能有可能被用来开发检测操纵性内容的新方法--部分原因是它具有丰富的大型基础模型的幻觉性能--从而帮助解决另一个现实世界的问题。

例如，1) 在健康主题中，LLM 和 VLM 代理的道德部署至关重要，尤其是在医疗保健等敏感领域。根据有偏见的数据训练出来的人工智能代理可能会为代表性不足的群体提供不准确的诊断，从而加剧健康差距。此外，人工智能代理处理敏感的患者数据也会引发重大的隐私和保密问题。2) 在游戏行业，人工智能代理可以改变开发人员的角色，将他们的工作重点从编写非玩家角色脚本转移到完善代理学习过程。同样，自适应机器人系统可以重新定义制造业的角色，需要新的技能组合，而不是取代人类工人。负责任地驾驭这些转变，对于最大限度地减少潜在的社会经济破坏至关重要。

此外，代理人工智能的重点是在模拟中学习协作政策，如果直接将政策应用于现实世界，则会因分布转移而存在一定风险。应建立健全的测试和持续的安全监控机制，以最大限度地降低真实世界场景中不可预测行为的风险。我们的 "VideoAnalytica" 数据集是从互联网上收集的，考虑到其不具有完全的代表性，因此我们已经通过了微软和华盛顿大学的伦理审查和法律程序。尽管如此，我们还需要了解该语料库中可能存在的偏差。数据分布的特征有很多。在本次研讨会中，我们已经捕捉到了我们数据集中的代理级分布与其他现有数据集的不同之处。然而，单个数据集或研讨会无法囊括的内容还有很多。我们认为，需要更多与实际任务或主题相关的方法或讨论，并通过提供这些数据或系统。

我们将在项目中专门讨论这些伦理问题，探索潜在的缓解策略，并部署一个负责任的多模式人工智能代理

。我们希望通过这篇论文帮助更多研究人员共同回答这些问题。

12 多样性声明

通过研究人工智能代理模型在不同领域的适应性，我们从本质上接受了排行榜、观点和解决方案的多样性。

因此，我们的项目旨在通过探索多模态和代理人工智能的广泛主题，建立一个多元化的社区。

本着这些原则，本项目重点关注在物理和虚拟环境中有效互动并促进与人类有效互动的先进多模态系统。因此，我们打算广泛邀请不同技术专业、文化、国家和学术领域的专家和从业人员参与讨论重要议题，包括但不限于：

- 基础模型的应用：开发具有综合模式（音频、图像、文本、传感器输入）的代理，旨在提高其在各种应用中的识别和响应能力。
- 通用端到端系统：开发使用大规模数据进行训练的端到端模型，力求创建多功能、适应性强的人工智能解决方案。
- 各种模式的接地方法：整合各种模式的信息，提高数据处理的一致性和效率。
- 直观的人类界面：开发人类与代理之间有效而有意义的互动。
- 驯服 LLM/VLM：探索解决大规模模型常见问题的新方法，如模型输出中的幻觉和偏差。

我们希望通过利用我们独特而多样的视角，拓宽我们对代理人工智能的潜力和局限性的集体理解。我们坚信，这种方法不仅能丰富个人观点，还能增强社区的集体知识，并促进形成一种更全面的视角，以包容多模态人工智能代理所面临的广泛挑战。

参考资料

M.M. Ahn、A. Brohan、N. Brown、Y. Chebotar、O. Cortes、B. David、C. Finn、C. Fu、K. Gopalakrishnan、K. Hausman、

A.Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. Joshi, R.Julian、D. Kalashnikov、Y. Kuang、K.-H.Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao

J.J. Rettinghouse、D. Reyes、P. Sermanet、N. Sievers、C. Tan、A. Toshev、V. Vanhoucke、F. Xia、T. Xiao、P. Xu、S. Xu,

M.Yan, and A. Zeng, "Do as I can and not as I say: 机器人能力中的语言基础》，载于 arXiv 预印本
arXiv:2204.01691, 2022。

M.M. Ahn、A. Brohan、N. Brown、Y. Chebotar、O. Cortes、B. David、C. Finn、K. Gopalakrishnan、K. Hausman、A. Herzog

et al., "Do as I can, not as I say: 机器人能力中的语言基础》，arXiv preprint arXiv:2204.01691, 2022。

J. -B.Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp.

A. Allevato, E. S. Short, M. Pryor, and A. Thomaz, "Tunenet: One-shot residual tuning for system identification and sim-to-real robot task transfer," in *Conference on Robot Learning*.PMLR, 2020, pp.

A. Amjad, P. Kordel, and G. Fernandes, "A review on innovation in healthcare sector (telehealth) through artificial intelligence," *Sustainability*, vol. 15, no. 8, p. 6655, 2023.

S.An, Z. Lin, Q. Fu, B. Chen, N. Zheng, J.-G. Lou, and D. Zhang, "How do in-context examples affect compositional generalization ? Lou, and D. Zhang, "How do in-context examples affect compositional generalization?" arXiv preprint arXiv:2305.04835, 2023.

P.Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva 等人, "论具身导航代理的评估", arXiv 预印本 arXiv:1807.06757, 2018。

P.Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: 解读真实环境中以视觉为基础的导航指令》，《电气与电子工程师协会计算机视觉与模式识别会议论文集》，2018 年，第 3674-3683 页。

S.Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *IEEE 计算机视觉国际会议论文集*, 2015 年, 第 2425-2433 页。

M.Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: 用于端到端检索的视频和图像联合编码器",《2021 年 IEEE/CVF 计算机视觉国际会议论文集》，第 1728-1738 页。

B. Baker, I. Akkaya, P. Zhokov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro, and J. Clune, "Video pretraining (vpt): 神经信息处理系统进展》，第 35 卷, 第 24 639-24 654 页, 2022 年。

- D.Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited : On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.
- K.Black, M. Nakamoto, P. Atreya, H. Walke, C. Finn, A. Kumar, and S. Levine, "Zero-shot robotic manipulation with pretrained image-editing diffusion models," *arXiv preprint arXiv:2310.10639*, 2023.
- A.Blair-Stanek, N. Holzenberger, and B. Van Durme, "Can gpt-3 perform statutory reasoning?" *arXiv preprint arXiv:2302.06100*, 2023.
- K.Bousmalis、G. Vezzani、D. Rao、C. Devin、A. X. Lee、M. Bauza、T. Davchev、Y. Zhou、A. Gupta、A. Raju 等人 , "Robocat : A self-improving foundation agent for robotic manipulation," *arXiv preprint arXiv:2306.11706*, 2023.
- A.Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu 等人, 《Rt-1：用于真实世界规模控制的机器人变压器》, *arXiv 预印本 arXiv:2212.06817*, 2022 年。
- A.Brohan、N. Brown、J. Carbajal、Y. Chebotar、X. Chen、K. Choromanski、T. Ding、D. Driess、A. Dubey、C. Finn 等人, "Rt-2：视觉-语言-动作模型将网络知识转化为机器人控制", *arXiv 预印本 arXiv:2307.15818*, 2023。

T.Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

S.Bubeck、V. Chandrasekaran、R. Eldan、J. Gehrke、E. Horvitz、E. Kamar、P. Lee、Y. T. Lee、Y. Li、S. Lundberg 等人, "人工通用智能的火花: 早期 gpt-4 实验", *arXiv 预印本 arXiv:2303.12712*, 2023。

W.Cai, S. Huang, G. Cheng, Y. Long, P. Gao, C. Sun, and H. Dong, "Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill," *arXiv preprint arXiv:2309.10309*, 2023.

M.Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan, "On the utility of learning about humans for human-ai coordination," *Advances in neural information processing systems*, vol. 32, 2019.

Y.Chang、M. Narang、H. Suzuki、G. Cao、J. Gao 和 Y. Bisk, "WebQA：Multihop and Multimodal QA," *arXiv preprint arXiv:2109.00590*, 2021.

D.Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems*, vol. 33, pp.

D.S. Chaplot、R. Salakhutdinov、A. Gupta 和 S. Gupta, "用于视觉导航的神经拓扑 Slam", 《2020 年 IEEE/CVF 计算机视觉与模式识别会议论文集》, 第 12 875-12 884 页。

G. Chen, Y. -D.Zheng, J. Wang, J. Xu, Y. Huang, J. Pan, Y. Wang, Y. Wang, Y. Qiao, T. Lu, and L. Wang, "Videollm：用大型语言模型对视频序列建模", 2023 年。

K.Chen, Q. Huang, H. Palangi, P. Smolensky, K. D. Forbus, and J. Gao, "Mapping natural-language problems to formal-language solutions using structured neural representation," in *ICML 2020*, July 2020.

K.Chen, Q . Huang, D . McDuff, X . Gao, H . Palangi, J . Wang, K . Forbus, and J. Gao, " Nice: Neural image commenting with empathy," in *EMNLP 2021*, October 2021.[在线]。Available: <https://www.microsoft.com/en-us/research/publication/nice-neural-image-commenting-with-empathy/>

J. H. Choi, K. E. Hickman, A. Monahan, and D. Schwarcz, "Chatgpt goes to law school," *Available at SSRN*, 2023.

H.H. W. Chung、L. Hou、S. Longpre、B. Zoph、Y. Tay、W. Fedus、Y. Li、X. Wang、M. Dehghani、S. Brahma 等人, "Scaling instruction-finetuned language models, " *arXiv preprint arXiv:2210.11416*, 2022。

N.N. C. F. Codella、D. Gutman、M. E. Celebi、B. Helba、M. A. Marchetti、S. W. Dusza、A. Kalloo、K. Liopyris、N. Mishra、

H.Kittler 和 A. Halpern, "面向黑色素瘤检测的皮肤病变分析: 国际皮肤成像合作组织 (ISIC) 主办的 2017 年国际生物医学成像研讨会 (ISBI) 上的挑战", 《2018 年第 15 届 IEEE 国际生物医学成像研讨会 (ISBI 2018) 》, 2018 年, 第 168-172 页。

A. Creswell, M. Shanahan, and I. Higgins, "Selection-inference: Exploiting large language models for interpretable logical reasoning," *arXiv preprint arXiv:2205.09712*, 2022.

B. Cui, A. Lupu, S. Sokota, H. Hu, D. J. Wu, and J. N. Foerster, "Adversarial diversity in hanabi," in *The Eleventh International Conference on Learning Representations*, 2023.[Online].Available: https://openreview.net/forum?id=uLE3WF3-H_5

G.Dagan、F.Keller 和 A.Lascarides, "用 llm 进行动态规划", *arXiv 预印本 arXiv:2308.06391*, 2023。

W.Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip：实现具有指令调整功能的通用视觉语言模型", 2023 年。

A. d'Avila Garcez 和 L. C. Lamb, "Neurosymbolic ai：The 3rd wave," 2020.

M.Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M.Wallingford 等人, "Robothor：开放式模拟到真实的人工智能平台", 《2020 年 IEEE/CVF 计算机视觉与模式识别会议论文集》, 第 3164-3174 页。

- Q.Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey for in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.
- V.V. S. Dorbala、G. Sigurdsson、R. Piramuthu、J. Thomason 和 G. S. Sukhatme, "Clip-nav：Using clip for zero-shot vision-and-language navigation," *arXiv preprint arXiv:2211.16649*, 2022.
- V.V. S. Dorbala, J. F. Mullen Jr, and D. Manocha, "Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation," *arXiv preprint arXiv:2303.03480*, 2023.
- A.Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold、S. Gelly 等人, "一幅图像胜过 16x16 个单词：大规模图像识别变换器"，*ICLR*, 2021 年。
- D.Driess、F. Xia、M. S. Sajjadi、C. Lynch、A. Chowdhery、B. Ichter、A. Wahid、J. Tompson、Q. Vuong、T. Yu 等人, "Palm-e：An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.
- Y.Du、M. Yang、P. Florence、F. Xia、A. Wahid、B. Ichter、P. Sermanet、T. Yu、P. Abbeel、J. B. Tenenbaum 等, "视频语言规划"，*arXiv 预印本 arXiv:2310.10625*, 2023。
- N.Dziri, A. Madotto, O. Zaiane, and A. J. Bose, "Neural path hunter：Reducing hallucination in dialogue systems via path grounding," *arXiv preprint arXiv:2104.08455*, 2021.
- K.Ehsani, W. Han, A. Herrasti, E. VanderBilt, L. Weihs, E. Kolve, A. Kembhavi, and R. Mottaghi, "Manipulathor：视觉对象操纵框架"，《2021 年 IEEE/CVF 计算机视觉与模式识别会议论文集》，第 4497-4506 页。
- D.Fried、R. Hu、V. Cirik、A. Rohrbach、J. Andreas、L.-P.Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and
T.达雷尔："用于视觉和语言导航的说话者跟随者模型"，《神经信息处理系统进展》(NIPS)，2018 年。
- T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu, "Violet : End-to-end video-language transformers with masked visual-token modeling," 2022.
- , "带遮蔽视觉建模的端到端视频语言转换器的实证研究"，2023 年。
- S.Y. Gadre、M. Wortsman、G. Ilharco、L. Schmidt 和 S. Song, "牧场上的奶牛：IEEE/CVF 计算机视觉与模式识别会议论文集"，2023 年，第 23 171-23 181 页。
- J.Gao, B. Peng, C. Li, J. Li, S. Shayandeh, L. Liden, and H.-Y. Shum, "Robust conversational ai with grounded text generation," *arXiv preprint arXiv:2009.03457*, 2020.Shum, "Robust conversational ai with grounded text generation," *arXiv preprint arXiv:2009.03457*, 2020.
- J.Gao, C. Xiong, P. Bennett, and N. Craswell, "Neural approaches to conversational information retrieval," *arXiv preprint arXiv:2201.05176*, 2022.
- C. R.Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, "Integrated task and motion planning," *Annual review of control, robotics, and autonomous systems*, vol. 4, pp.
- T.Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," *Science Robotics*, vol. 8, no. 79, p. eadf6991, 2023.
- R.Gong, J. Huang, Y. Zhao, H. Geng, X. Gao, Q. Wu, W. Ai, Z. Zhou, D. Terzopoulos, S.-C.Zhu 等人，"Arnold：

现实 3D 场景中具有连续状态的语言基础任务学习基准", 《IEEE/CVF 计算机视觉国际会议 (ICCV) 论文集》, 2023 年。

R.Gong, Q. Huang, X. Ma, H. Vo, Z. Durante, Y. Noda, Z. Zheng, S.-C. Zhu, D. Terzopoulos, L. Fei-Fei et al.Zhu、D. Terzopoulos 、 L. Fei-Fei 等人 , "Mindagent : Emergent gaming interaction," *arXiv preprint arXiv:2309.09971*, 2023.

A.Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation,".Chao 和 D. Fox, "Rvt: 用于三维物体操作的机器人视图转换器"。
arXiv preprint arXiv:2306.14896, 2023.

M.Gramopadhye 和 D. Szafir, "利用环境感知语言模型生成可执行的行动计划"。

arXiv preprint arXiv:2210.04964, 2022.

A.Gudibande, E. Wallace, C. Snell, X. Geng, H. Liu, P. Abbeel, S. Levine, and D. Song, "The false promise of imitating proprietary llms," *arXiv preprint arXiv:2305.15717*, 2023.

L.Gui, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "Vlc：从字幕训练视觉语言转换器", 2022 年 5 月。

L.Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "Kat: A knowledge augmented transformer for vision-and-language," in *NAACL 2022. 长篇论文, 口述*, 2022 年 1 月。

R.R. L. Guimarães、A. S. de Oliveira、J. A. Fabro、T. Becker 和 V. A. Brenner, "Ros 导航：概念与教程"。

机器人操作系统 (ROS) 完全参考》(第1卷)，第 121-160 页, 2016 年。

K.Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International conference on machine learning*.PMLR, 2020, pp.

H.Ha, P. Florence, and S. Song, "Scaling up and distilling down：语言引导的机器人技能习得》，*arXiv preprint arXiv:2307.14535*, 2023.

T.Haarnoja 、 B. Moran 、 G. Lever 、 S. H. Huang 、 D. Tirumala 、 M. Wulfmeier 、 J. Humplik 、 S. Tunyasuvunakool、 N. Y. Siegel、 R. Hafner 等人, "利用深度强化学习为双足机器人学习敏捷足球技能", *arXiv 预印本 arXiv:2304.13653*, 2023。

K.He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *CVPR*, 2022.

S.Hemachandra, F. Duvallet, T. M. Howard, N. Roy, A. Stentz, and M. R. Walter, "Learning models for following natural language directions in unknown environments," *arXiv preprint arXiv:1503.05079*, 2015.

J. Henrich、S. J. Heine 和 A. Norenzayan, "世界上最奇怪的人？行为与脑科学》，第 33 卷, 第 2-3 期, 第 61-83 页, 2010 年。

L.L. B. Hensel, N. Yongsatianchot, P. Torshizi, E. Minucci, and S. Marsella, "Large language models in textual analysis for gesture selection," in *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, 2023, pp.

D. Ho, K. Rao, Z. Xu, E. Jang, M. Khansari, and Y. Bai, "Retinagan：从模拟到现实的物体感知方法", 2021 年 *IEEE 机器人与自动化国际会议 (ICRA)* , IEEE, 2021 年, 第 10 920-10 926 页。 IEEE, 2021, pp.

C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*.IEEE, 2023, pp.

Q.Huang、J. S. Park、A. Gupta、P. Bennett、R. Gong、S. Som、B. Peng、O. K. Mohammed、C. Pal、Y. Choi 等人, "Ark：具有知识交互新兴能力的增强现实", *arXiv 预印本 arXiv:2305.00970*, 2023。

W.Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners：Extracting actionable knowledge for embodied agents," in *Proceedings of the 39th International Conference on Machine Learning*, ser. *机器学习研究论文集》* , K. Chaudhuri、S. Jegelka、L. Song、C. Szepesvari,

G. Niu 和 S. Sabato 编辑，第 162 卷。PMLR, 17-23 Jul 2022, pp.[Online].Available:
<https://proceedings.mlr.press/v162/huang22a.html>

W.Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet
、

N.Brown, T. Jackson, L. Luu, S. Levine, K. Hausman, and B. Ichter, "Inner monologue：通过语言模型规划体
现推理"，载于 *arXiv 预印本 arXiv:2207.05608*, 2022。

Z.Huang, H. Feng, Z. Chongzhi, L. Sheng, L. Ziwei, and J. Shao, "Dolphin：基于 llms 的通用视频交互平台"，
2023 年，<https://github.com/kaleido-lab/dolphin>。

K.Ikeuchi, N. Wake, K. Sasabuchi, and J. Takamatsu, "Semantic constraints to represent common sense required in household actions for multimodal learning-from-observation robot," *The International Journal of Robotics Research*, vol. 0, no.0, p. 02783649231212929, 0.

K.Ikeuchi, J. Takamatsu, K. Sasabuchi, N. Wake, and A. Kanehiro, "Applying learning-from-observation to household service robots: three common-sense formulation," *arXiv preprint arXiv:2304.09966*, 2023.

S.Imani, L. Du, and H. Shrivastava, "Mathprompter: Mathematical reasoning using large language models," *arXiv preprint arXiv:2303.05398*, 2023.

S.James and A. J. Davison, "Q-attention: Enabling efficient learning for vision-based robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp.

E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*.PMLR, 2022, pp.

Z.Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp.

Y.Jiang、A. Gupta、Z. Zhang、G. Wang、Y. Dou、Y. Chen、L. Fei-Fei、A. Anandkumar、Y. Zhu 和 L. Fan, "Vima: 多模态提示的通用机器人操纵", *arXiv*, 2022 年。

D. Kalashnikov、A. Irpan、P. Pastor、J. Ibarz、A. Herzog、E. Jang、D. Quillen、E. Holly、M. Kalakrishnan、V. Vanhoucke 等人, "基于视觉的机器人操纵的可扩展深度强化学习", *机器人学习大会*。PMLR, 2018 年, 第 651-673 页。

A.Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," (双向图像句子映射的深度片段嵌入)

神经信息处理系统进展 , 第 27 卷, 2014 年。

P.Katara, Z. Xian, and K. Fragkiadaki, "Gen2sim: Scaling up robot learning in simulation with generative models," (《Gen2sim: 利用生成模型在仿真中扩大机器人学习规模》)。
arXiv preprint arXiv:2310.18308, 2023.

L.Ke, X. Li, B. Yonatan, A. Holtzman, Z. Gan, J. Liu, J. Gao, Y. Choi, and S. Srinivasa, "Tactical rewind: 通过视觉和语言导航中的回溯进行自我校正", 《电气和电子工程师协会计算机视觉与模式识别大会 (CVPR) 论文集》, 2019 年。

J.Kim, J. Kim, and S. Choi, "Flame: 基于自由格式语言的动作合成与编辑", 《AAAI 人工智能会议论文集》, 第 37 卷, 第 7 期, 2023 年, 第 8255-8263 页。

P.-C. Ko, J. Mao, Y. Du, S.-H.Ko, J. Mao, Y. Du, S. -H.Sun, and J. B. Tenenbaum, "Learning to act from actionless videos through dense correspondences," *arXiv preprint arXiv:2310.08576*, 2023.

E. Kolve、R. Mottaghi、W. Han、E. VanderBilt、L. Weihs、A. Herrasti、M. Deitke、K. Ehsani、D. Gordon、Y. Zhu 等人, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

R.Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma、

M.Bernstein 和 L. Fei-Fei, "Visual genome：利用众包密集图像注释连接语言和视觉》，载于
arXiv:1602.07332, 2016。

K.N. Kumar, I. Essa, and S. Ha, "Words into action：使用语言引导的迭代动作精炼学习多样化仿人机器人行为
》，*arXiv 预印本 arXiv:2310.06226*, 2023.

P.Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine," *New England Journal of Medicine*, vol. 388, no. 13, pp.

P.Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t.Yih, T. Rocktäschel *et al.*,
"Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp.

B.Li, P. Wu, P. Abbeel, and J. Malik, "Interactive task planning with language models," *arXiv preprint arXiv:2310.10645*,
2023.

C.Li、F. Xia、R. Martín-Martín、M. Lingelbach、S. Srivastava、B. Shen、K. Vainio、C. Gokmen、G. Dharan、T. Jain 等人, "igibson 2.0: 以对象为中心的机器人日常家务学习模拟", *arXiv 预印本 arXiv:2108.03272*, 2021。

C.Li、C. Wong、S. Zhang、N. Usuyama、H. Liu、J. Yang、T. Naumann、H. Poon 和 J. Gao, "Llava-med: 一天内训练出一个大型生物医学语言和视觉助手》, *arXiv 预印本 arXiv:2306.00890*, 2023。

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: 大型语言模型社会的 "心智" 探索交流代理》, *arXiv 预印本 arXiv:2303.17760*, 2023。

J.Li, Q. Gao, M. Johnston, X. Gao, X. He, S. Shakiah, H. Shi, R. Ghanadan, and W. Y. Wang, "Mastering robot manipulation with multimodal prompts through pretraining and multi-task fine-tuning," *arXiv preprint arXiv:2310.09676*, 2023.

J.Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

K.K.Li、Y. He、W. Yi、Y. Li、W. Wang、P. Luo、Y. Wang、L. Wang 和 Y. Qiao, "Videochat: Video Chat : 视频聊天"。Qiao, "Videochat: 以聊天为中心的视频理解》, *arXiv 预印本 arXiv:2305.06355*, 2023。

L.Li、Y.-C.Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "Hero: Hierarchical encoder for video+language omni-representation pre-training," 2020.

L.Li、J. Lei、Z. Gan、L. Yu、Y.-C.Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang, T. L. Berg, M. Bansal,

J.Liu, L. Wang, and Z. Liu, "Value: 用于视频和语言理解评估的多任务基准", 2021 年。

X.Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu *et al.*

J.Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: 体现控制的语言模型程序", 载于 *arXiv 预印本 arXiv:2209.07753*, 2022。

X.Liang、L. Ma、S. Guo、J. Han、H. Xu、S. Ma 和 X. Liang, "Mo-vln: A multi-task benchmark for open-set zero-shot vision-and-language navigation," *arXiv preprint arXiv:2306.10322*, 2023.

S.Lifshitz, K. Paster, H. Chan, J. Ba, and S. McIlraith, "Steve-1: A generative model for text-to-behavior in minecraft," *arXiv preprint arXiv:2306.00937*, 2023.

K.Lin、F. Ahmed、L. Li、C.-C.Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu, C. Liu, and L. Wang, "Mm-vid: Advancing video understanding with gpt-4v(ision)," 2023.

T.-Y.林Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: 上下文中的通用对象", 《2014 年 ECCV 论文集》。

C.C. K. Liu 和 D. Negrut, "基于物理的模拟器在机器人学中的作用", 《控制、机器人学和自主系统年度评

论》，第 4 卷，第 35-58 页，2021 年。

H.Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

H.Liu, A. Chen, Y. Zhu, A. Swaminathan, A. Kolobov, and C.-A. Cheng, "Interactive robot learning from verbal correction," arXiv preprint arXiv:2310.17555, 2023.Cheng, "交互式机器人从口头纠正中学习", *arXiv preprint arXiv:2310.17555*, 2023.

Y.Liu, W. Held, and D. Yang, "Dada：通过语言规则的动态聚合实现方言适应", 《2023 年自然语言处理实证方法会议论文集》, 2023 年。

P.Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S. -C.Chang, Y. N. Wu, S.-C. Zhu, and J. Gao, "Chameleon: Chameleon.Zhu, and J. Gao, "Chameleon：大型语言模型的即插即用组合推理", 2023 年。

Z.Luo、Z. Durante、L. Li、W. Xie、R. Liu、E. Jin、Z. Huang、L. Y. Li、J. Wu、J. C. Niebles 等人, "Moma-lrg: 神经信息处理系统进展》, 第 35 卷, 第 5282-5298 页, 2022 年。

C.-Y.Ma, J. Lu, Z. Wu, G. AlRegib, Z. Kira, R. Socher, and C. Xiong, "Self-monitoring navigation agent via auxiliary progress estimation," *arXiv preprint arXiv:1901.03035*, 2019.

C.-Y.Ma, Z. Wu, G. AlRegib, C. Xiong, and Z. Kira, "The regretful agent: Heuristic-aided navigation through progress estimation," *arXiv preprint arXiv:1903.01602*, 2019.

Y.Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," Oxford University Press.Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," *arXiv preprint arXiv:2310.12931*, 2023.

M.Maaz、H. Rasheed、S. Khan 和 F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," 2023.

R.Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: 基于提示的情感分析和情绪检测的实证研究》, 《电气和电子工程师学会情感计算期刊》, 2022 年。

G. Marcus, "The next decade in ai: four steps towards robust artificial intelligence," *arXiv preprint arXiv:2002.06177*, 2020.

G. Marcus 和 E. Davis, 《重启人工智能：打造值得信赖的人工智能》。 Pantheon, 2019.

K.Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: 需要外部知识的可视化问题解答基准", *CVPR* , 2019。

P.Martinez-Gonzalez, S. Oprea, A. Garcia-Garcia, A. Jover-Alvarez, S. Orts-Escalano, and J. Garcia-Rodriguez, "Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation," *Virtual Reality*, vol. 24, pp.

J.J. Maynez、S. Narayan、B. Bohnet 和 R. McDonald, "论抽象概括中的忠实性和事实性", 载《美国心理学报》 , 2011 年。

第 58 届计算语言学协会年会论文集》, D. Jurafsky, J. Chai、
N.Schluter, and J. Tetreault, Eds.Online: 计算语言学协会, 2020 年 7 月, 第 1906-1919 页。
[Online].Available: <https://aclanthology.org/2020.acl-main.173>

O.Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for longhorizon *robot* manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no.3, pp.

Meta Fundamental AI Research (FAIR) Diplomacy Team, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty、D. 弗里德、A. 戈夫、J. 格雷、H. 胡等人, "通过将语言模型与战略推理相结合在外交游戏中进行人类水平的游戏", 《科学》, 第 378 卷, 第 6624 期, 第 1067-1074 页, 2022 年。

S.Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: *arXiv preprint arXiv:2202.12837*, 2022.

M.I. Minsky, "Minsky's frame system theory," in *Proceedings of the 1975 Workshop on Theoretical Issues in*

Natural Language Processing, ser.TINLAP '75. 美国：计算语言学协会，1975年，第104-116页。
[Online]. Available: <https://doi.org/10.3115/980190.980222>

S.Mirchandani、F. Xia、P. Florence、B. Ichter、D. Driess、M. G. Arenas、K. Rao、D. Sadigh 和 A. Zeng, 《作为通用模式机器的大型语言模型》, *arXiv 预印本 arXiv:2307.04721*, 2023。

P.P. Mirowski、R. Pascanu、F. Viola、H. Soyer、A. J. Ballard、A. Banino、M. Denil、R. Goroshin、L. Sifre、K. Kavukcuoglu

等人, 《在复杂环境中学习导航》, *arXiv 预印本 arXiv:1611.03673*, 2016。

M.Mittal、C. Yu、Q. Yu、J. Liu、N. Rudin、D. Hoeller、J. L. Yuan、R. Singh、Y. Guo、H. Mazhar 等人, "Orbit: 交互式机器人学习环境的统一仿真框架", 《电气与电子工程师学会机器人与自动化通讯》, 2023年。

V.Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski 等人, "通过深度强化学习实现人级控制", 《自然》, 第 518 卷, 第 7540 期, 第 529-533 页, 2015 年。

A.Mousavian, A. Toshev, M. Fiser, J. Kosecka, and J. Davidson, "Visual representations for semantic target driven navigation," *arXiv preprint arXiv:1805.06066*, 2018.

T.Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: 大规模演示的通用操作技能基准", *arXiv 预印本 arXiv:2107.14483*, 2021。

M.Müller, V. Casser, J. Lahoud, N. Smith, and B. Ghanem, "Sim4cv: 用于计算机视觉应用的照片逼真模拟器", 《国际计算机视觉杂志》, 第 126 卷, 第 902-919 页, 2018 年。

J.J. Nay, "Law informs code: A legal informatics approach to aligning artificial intelligence with humans," *Nw.J. Tech.Prop.*, 第 20 卷, 第 309 页, 2022 年。

K.Nguyen, D. Dey, C. Brockett, and B. Dolan, "Vision-based navigation with language-based assistance via imitation learning with indirect intervention," *arXiv preprint arXiv:1812.04155*, 2018.

Z.Ni, X.-X. Deng, C. Tai, X.-Y.Deng, C. Tai, X.-Y.Zhu, X. Wu, Y.-J. Liu, and L. Zeng, "Grid: 基于场景图的指令驱动型机器人任务规划", *arXiv 预印本 arXiv:2309.07726*, 2023.

OpenAI, "GPT-4 技术报告", OpenAI, Tech.Rep., 2023.

L.L. Ouyang、J. Wu、X. Jiang、D. Almeida、C. Wainwright、P. Mishkin、C. Zhang、S. Agarwal、K. Slama、A. Ray 等人, "训练语言模型以遵循人类反馈指令", 《神经信息处理系统进展》, 第 35 卷, 第 27 730-27 744 页, 2022 年。

A.Padalkar、A. Pooley、A. Jain、A. Bewley、A. Herzog、A. Irpan、A. Khazatsky、A. Rai、A. Singh、A. Brohan 等人, "开放 x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

M.Parakh, A. Fong, A. Simeonov, A. Gupta, T. Chen, and P. Agrawal, "Human-assisted continual robot learning with foundation models," *arXiv preprint arXiv:2309.14321*, 2023.

J.Park, J. Hessel, K. Chandu, P. P. Liang, X. Lu, P. West, Q. Huang, J. Gao, A. Farhadi, and Y. Choi, "Multimodal agent - localized symbolic knowledge distillation for visual commonsense models," in *NeurIPS 2023*, October 2023.

J.Park, J. Hessel, K. Chandu, P. P. Liang, X. Lu, P. West, Y. Yu, Q. Huang, J. Gao, A. Farhadi, and Y. Choi, "Localized symbolic knowledge distillation for visual commonsense models," in *Thirty-seventh Conference on Neural Information Processing Systems, 2023*.[Online].Available: <https://openreview.net/forum?id=V5eG47pyVI>

J.Park, Q. Huang, Y. Bisk, J. Yang, S. Som, A. Farhadi, Y. Choi, and J. Gao, "Ink: Intensive neural knowledge," July 2022.

J.J. S. Park、J. C. O'Brien、C. J. Cai、M. R. Morris、P. Liang 和 M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," *arXiv preprint arXiv:2304.03442*, 2023.

B.Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen *et al*: 利用外部知识和

自动反馈改进大型语言模型》，*arXiv 预印本 arXiv:2302.12813, 2023。*

X.X. B. Peng、M. Andrychowicz、W. Zaremba 和 P. Abbeel，"带动力学随机化的机器人控制的仿真到真实传输"，*2018 IEEE 机器人与自动化国际会议 (ICRA)*。IEEE, 2018, pp.

X.Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, "Virtualhome：通过程序模拟家庭活动"，*2018 年 IEEE 计算机视觉与模式识别国际会议 (CVPR)*，2018 年，第 8494-8502 页。

X.Puig, E. Undersander, A. Szot, M. D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A. W. Clegg, M. Hlavac, S. Y. Min.Yang, R. Partsey, R. Desai, A. W. Clegg, M. Hlavac, S. Y. Min

等人，"栖息地 3.0：人类、化身和机器人的共同栖息地"，*arXiv 预印本 arXiv:2310.13724, 2023。*

A.Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark 等人, "从自然语言监督中学习可转移的视觉模型", *机器学习国际会议*。PMLR, 2021, pp.

S.K. Ramakrishnan、D. S. Chaplot、Z. Al-Halah、J. Malik 和 K. Grauman, "Poni: Poni: Potential functions for objectgoal navigation with interaction-free learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp.

S.S. S. Raman、V. Cohen、D. Paulius、I. Idrees、E. Rosen、R. Mooney 和 S. Tellex, "Cape: 使用大型语言模型从前提条件错误中纠正行动", *第二届语言与机器人学习研讨会：语言作为基础* , 2023 年。

K.Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari, "Rl-cyclegan: 2020年IEEE/CVF计算机视觉与模式识别会议论文集》、
pp.11 157-11 166.

V.Raunak、A. Menezes 和 M. Junczys-Dowmunt, "神经机器翻译中幻觉的奇特案例"。
arXiv preprint arXiv:2104.06683, 2021.

S.Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg et al., "A generalist agent," *arXiv preprint arXiv:2205.06175*, 2022.

M.Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," *Advances in neural information processing systems*, vol. 28, 2015.

A.Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," *arXiv preprint arXiv:1809.02156*, 2018.

A.Rosinol, J. J. Leonard, and L. Carlone, "Nerf-slam: 利用神经辐射场的实时密集单目斜视"。
arXiv preprint arXiv:2210.13641, 2022.

F.Sadeghi 和 S. Levine, "Cad2rl: 无单个真实图像的真实单图像飞行, " *arXiv 预印本 arXiv:1611.04201*, 2016
◦

D.Saito, K. Sasabuchi, N. Wake, J. Takamatsu, H. Koike, and K. Ikeuchi, "Task-grasping from a demonstrated human strategy," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, 2022, pp.

D.Saito, K. Sasabuchi, N. Wake, A. Kanehira, J. Takamatsu, H. Koike, and K. Ikeuchi, "Constraint-aware policy for compliant manipulation," 2023.

B.Sarkar, A. Shih, and D. Sadigh, "Diverse conventions for human-AI collaboration," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

K.Sasabuchi, N. Wake, and K. Ikeuchi, "Task-oriented motion mapping on robots of various configuration using body role division," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp.

K.Sasabuchi, D. Saito, A. Kanehira, N. Wake, J. Takamatsu, and K. Ikeuchi, "Task-sequencing simulator: 集成机器学习到机器人操纵的执行模拟, " *arXiv 预印本 arXiv:2301.01382*, 2023.

M.Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun, "Minos: 用于复杂环境导航的多模态室内模拟器》, *arXiv 预印本 arXiv:1712.03931*, 2017。

M.Savva、A. Kadian、O. Maksymets、Y. Zhao、E. Wijmans、B. Jain、J. Straub、J. Liu、V. Koltun、J. Malik 等

λ, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp.

T.T.Schick、J.Dwivedi-Yu、R.Dessì、R.Raileanu、M.Lomeli、L.Zettlemoyer、N.Cancedda 和 T.Scialom , "Toolformer: 语言模型可以自学使用工具》, 2023 年。

E.Segalis, D. Valevski, D. Lumen, Y. Matias, and Y. Leviathan, "A picture is worth a thousand words: Principled recaptioning improves image generation," *arXiv preprint arXiv:2310.16656*, 2023.

B.Shacklett, L. G. Rosenzweig, Z. Xie, B. Sarkar, A. Szot, E. Wijmans, V. Koltun, D. Batra, and K. Fatahalian, "An extensible, data-oriented architecture for high-performance, many-world simulation," *ACM Trans. Graph.* 4, 2023.

D.Shah、B. Osin'ski、S. Levine 等人, "Lm-nav：使用预先训练的大型语言、视觉和动作模型进行机器人导航", 机器人学习大会。PMLR, 2023, pp.

R.Shah, R. Martín-Martín, and Y. Zhu, "Mutex : Learning unified policies from multimodal task specifications," *arXiv preprint arXiv:2309.14320*, 2023.

S.Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics : 第11届国际会议成果*。Springer, 2018, pp.

P.Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions：A cleaned, hypernymed, image alt-text dataset for automatic image captioning," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

R.Shi, Y. Liu, Y. Ze, S. S. Du, and H. Xu, "Unleashing power of pre-trained language models for offline reinforcement learning," *arXiv preprint arXiv:2310.20587*, 2023.

M.Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor：用于机器人操纵的多任务转换器", 载于 机器人学习大会。 PMLR, 2023 年, 第 785-799 页。

K.Shuster、S. Poff、M. Chen、D. Kiela 和 J. Weston, "检索增强减少了对话中的幻觉"。
arXiv preprint arXiv:2104.07567, 2021.

A.Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp.

S.Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

S.S. A. Sontakke, J. Zhang, S. M. Arnold, K. Pertsch, E. Büyükköy, D. Sadigh, C. Finn, and L. Itti, "Roboclip：一次演示足以学习机器人策略", *arXiv preprint arXiv:2310.07899*, 2023.

S.Srivastava、C. Li、M. Lingelbach、R. Martín-Martín、F. Xia、K. E. Vainio、Z. Lian、C. Gokmen、S. Buch 、K. Liu 等人, "行为：虚拟、交互和生态环境中日常家务活动的基准", 机器人学习大会。PMLR, 2022, pp.

J.Sun, Q. Zhang, Y. Duan, X. Jiang, C. Cheng, and R. Xu, "Prompt, plan, perform：通过量化模仿学习实现基于 Llm 的仿人控制", *arXiv 预印本 arXiv:2309.11359*, 2023.

Q.Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip：改进的规模剪辑训练技术》, *arXiv 预印本 arXiv:2303.15389*, 2023。

A.Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets、A.Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D.Batra, "Habitat 2.0：训练家庭助理重新布置它们的栖息地》, 《神经信息处理系统进展》(NeurIPS) ,

2021 年。

J.Takamatsu, K. Sasabuchi, N. Wake, A. Kanehira, and K. Ikeuchi, "Learning-from-observation system considering hardware-level reusability," *arXiv preprint arXiv:2212.09242*, 2022.

C.Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, "GraspGPT: 利用大型语言模型中的语义知识进行任务导向抓取", 《电气与电子工程师学会机器人与自动化通讯》, 2023 年。

R.Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: 一种指令跟随骆驼模型", https://github.com/tatsu-lab/stanford_alpaca, 2023。

H.Teshima, N. Wake, D. Thomas, Y. Nakashima, H. Kawasaki, and K. Ikeuchi, "Deep gesture generation for social robots using type-specific libaries," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.IEEE, 2022, pp.

G. Tevet、S. Raab、B. Gordon、Y. Shafir、D. Cohen-Or 和 A. H. Bermano, "人体运动扩散模型", *arXiv 预印本 arXiv: 2209.14916*, 2022 年。

J.Thomason, D. Gordan, and Y. Bisk, "Shifting the baseline: 视觉导航和 QA 的单一模式性能"。
arXiv preprint arXiv:1811.00613, 2018.

J.Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*.IEEE, 2017, pp.

H.H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. -A.Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar
等人, "Llama: 开放而高效的基础语言模型", *arXiv 预印本 arXiv:2302.13971*, 2023。

N.Tsoi, A. Xiang, P. Yu, S. S. Sohn, G. Schwartz, S. Ramesh, M. Hussein, A. W. Gupta, M. Kapadia, and M. Vázquez, "Sean 2.0 : Formalizing and generating social situations for robot navigation," *IEEE Robotics and Automation Letters*, vol. 7, no.4, pp.

N.Wake, R. Arakawa, I. Yanokura, T. Kiyokawa, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "A learning-from observation framework：用于抓取-操纵-释放家庭操作的一次性机器人教学", *2021 年 IEEE/SICE 系统集成国际研讨会 (SII)* 。 IEEE, 2021.

N.Wake, I. Yanokura, K. Sasabuchi, and K. Ikeuchi, "Verbal focus-of-attention system for learning-from-observation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp.

N.Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "Interactive task encoding system for learning-from- observation," in *2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, 2023, pp.

--, "Bias in emotion recognition with chatgpt, " *arXiv preprint arXiv:2310.11753*, 2023。

--, "Gpt 模型满足机器人应用：Co-speech gesturing chat system," *arXiv preprint arXiv:2306.01741*, 2023.

--, "Gpt-4v(ision) for robotics： 来自人类演示的多模态任务规划》, *arXiv 预印本 arXiv:2311.12015*, 2023。

--, "Chatgpt 在各种环境中的长步机器人控制：A case application," *IEEE Access*, vol. 11、
pp.95 060-95 078, 2023.

N.Wake, D. Saito, K. Sasabuchi, H. Koike, and K. Ikeuchi, "Text-driven object affordance for guiding grasp-type recognition in multimodal robot teaching," *Machine Vision and Applications*, vol. 34, no.4, p. 58, 2023.

B.Wang, Q. Huang, B. Deb, A. L. Halfaker, L. Shao, D. McDuff, A. Awadallah, D. Radev, and J. Gao, "Logical transformers：将逻辑结构注入预训练语言模型", *ACL 2023 年论文集*, 2023 年 5 月。

D.Wang, Q. Huang, M. Jackson, and J. Gao, "Retrieve what you need：开放域问题解答的相互学习框架", 2023 年 3 月。[在线]。 Available: <https://www.microsoft.com/en-us/research/publication/retrieve-what-you-need-a-mutual-learning-framework-for-open-domain-question-answering/>.

G. Wang、Y. Xie、Y. Jiang、A. Mandlekar、C. Xiao、Y. Zhu、L. Fan 和 A. Anandkumar, "Voyager：具有大型语言模型的开放式嵌入式代理", *arXiv 预印本 arXiv:2305.16291*, 2023。

L.Wang、C. Ma、X. Feng、Z. Zhang、H. Yang、J. Zhang、Z. Chen、J. Tang、X. Chen、Y. Lin 等: 《基于大型语言模型的自主代理调查》, *arXiv 预印本 arXiv:2308.11432*, 2023。

P.Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," *arXiv preprint arXiv:1511.02570*, 2015.

P.Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Hengel, "Fvqa：基于事实的可视化问题解答》，《TPAMI》，第 40 卷，第 10 期，第 2413-2427 页，2017 年。

X.Wang, W. Xiong, H. Wang, and W. Y. Wang, "Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation," in *The European Conference on Computer Vision (ECCV)*, September 2018.

X.Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Weng, W. Y. Wang and L. Zhang, in CVPR 2019 June 2019.Weng, W. Y. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in *CVPR 2019*, June 2019.

Y.Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Chen, Y. Wang, P. Luo, Z. Liu, Y. Wang, L. Wang, and Y. Qiao, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," 2023.Qiao, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," 2023.

Y.Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct：对齐语言模型与自生成指令》，*arXiv 预印本 arXiv:2212.10560*, 2022。

Y.Wang, Z. Xian, F. Chen, T.-H. Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held, and C. Gan, "Robogen: Robogen.Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held, and C. Gan, "Robogen：Towards unleashing infinite data for automated robot learning via generative simulation," *arXiv preprint arXiv:2311.01455*, 2023.

Z.Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang, "Humanise: Language-conditioned h u m a n motion generation in 3d scenes," in *Advances in Neural Information Processing Systems*, S. Koyejo, S.Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp.[Online].Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/6030db5195150ac86d942186f4abdad8-Paper-Conference.pdf

Z.Wang, S. Cai, A. Liu, X. Ma, and Y. Liang, "Describe, explain, plan and select：使用大型语言模型进行交互式规划，实现开放世界多任务代理》，*arXiv 预印本 arXiv:2302.01560*, 2023。

J.Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp.

世界卫生组织和世界银行，"追踪全民医保：第一份全球监测报告》，
www.who.int/healthinfo/universal_health_coverage/report/2015/en, 2015 年 6 月。

Q.Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang, "Autogen：Autogen: Enabling next-gen llm applications via multi-agent conversation," Microsoft, Tech.MSR-TR-2023-33, 2023 年 8 月。[在线]。 Available: <https://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-llm-applications-via-multi-agent-conversation-framework/>

Z.Xi、W. Chen、X. Guo、W. He、Y. Ding、B. Hong、M. Zhang、J. Wang、S. Jin、E. Zhou *et al.*：A survey," *arXiv preprint arXiv:2309.07864*, 2023.

F.Xia, A. R. Zamir, Z. -Y.He, A. Sax, J. Malik, and S. Savarese, "Gibson Env: real-world perception for embodied agents," in *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*.IEEE, 2018.

M.Xu, P. Huang, W. Yu, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, and D. Zhao, "Creative robot tool use

with large language models," *arXiv preprint arXiv:2310.13065*, 2023.

J.Yang, Y. Dong, S. Liu, B. Li, Z. Wang, C. Jiang, H. Tan, J. Kang, Y. Zhang, K. Zhou *et al*: 来自环境反馈的嵌入式视觉语言编程器", *arXiv preprint arXiv:2310.08588*, 2023.

K.Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis," *arXiv preprint arXiv:2304.03347*, 2023.

Z.Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang, "Mm-react : Prompting chatgpt for multimodal reasoning and action," 2023.

S.Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts：用大型语言模型故意解决问题", 2023 年。

S.Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React：语言模型中推理与行动的协同", 2023 年。

Q.叶 H. 徐 G. 徐 J. 叶 M. 严 Y. 周 J. 王 A. 胡 P. 史 Y. 史 C. 李 Y. 徐 H. 陈 J. 田 Q. 廉
J.Zhang 和 F. Huang, "mplug-owl：模块化增强了多模态大型语言模型的能力", 2023 年。

Y.Ye, H. You, and J. Du, "Improved trust in human-robot collaboration with chatgpt," *IEEE Access*, 2023.

P.Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations：事件描述语义推理的新相似度量", 《计算语言学协会年会论文集》, 2014 年。

J.Yu、X. Wang、S. Tu、S. Cao、D. Zhang-Li、X. Lv、H. Peng、Z. Yao、X. Zhang、H. Li 等人, "Kola：
Carefully benchmarking world knowledge of large language models," *arXiv preprint arXiv:2306.09296*, 2023.

L.Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Computer Vision-ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II* 14. Springer, 2016, pp.

W.Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T.Lee, M. G. Arenas, H.-T.L. Chiang, T. Erez, L. Hasenclever, J. Humplik
等人, 《从语言到奖励的机器人技能合成》, *arXiv 预印本 arXiv:2306.08647*, 2023。

R.Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot：多模态神经脚本知识模型", 2021 年。

R.Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi, "Merlot reserve
：通过视觉、语言和声音的神经脚本知识", 2022 年。

A.Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V.
Sindhwani et al：重新安排机器人操纵的视觉世界", 机器人学习大会。PMLR, 2021, pp.

A.Zeng, M. Liu, R. Lu, B. Wang, X. Liu, Y. Dong, and J. Tang, "Agenttuning：Enabling generalized agent abilities
for llms," 2023.

L.Zha, Y. Cui, L.-H.Lin, M. Kwon, M. G. Arenas, A. Zeng, F. Xia, and D. Sadigh, "Distilling and retrieving
generalizable knowledge for robot manipulation via language corrections," *arXiv preprint arXiv:2311.10678*,
2023.

M.Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation
with diffusion model," *arXiv preprint arXiv:2208.15001*, 2022.

S.Zhang, X. Song, Y. Bai, W. Li, Y. Chu, and S. Jiang, "Hierarchical object-to-zone graph for object navigation," in
the United Nations Journal of Technology, 2008.

IEEE/CVF 计算机视觉国际会议论文集, 2021 年, 第 15130-15140 页。

W.Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is chatgpt equipped with emotional dialogue capabilities?"
arXiv preprint arXiv:2304.09582, 2023.

Y.Zhao、I. Misra、P. Krähenbühl 和 R. Girdhar, "从大型语言模型中学习视频表征", 《语言学报》, 2011 年

◦

arXiv preprint arXiv:2212.04501, 2022.

L.Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023.

Z.Zhong、J. Cao、S. Gu、S. Xie、W. Gao、L. Luo、Z. Yan、H. Zhao 和 G. Zhou, "Assist: 用于可扩展的逼真室内模拟的交互式场景节点", *arXiv 预印本 arXiv:2311.06211*, 2023。

G. Zhou, Y. Hong, and Q. Wu, "Navgpt: 使用大型语言模型进行视觉语言导航的显式推理", *arXiv 预印本 arXiv:2305.16986*, 2023。

- H.Zhou, M. Ding, W. Peng, M. Tomizuka, L. Shao, and C. Gan, "Generalizable longhorizon manipulations with large language models," *arXiv preprint arXiv:2310.02264*, 2023.
- X.Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *ECCV*, 2022.
- Y.Zhou, C. Cui, J. Yoon, L. Zhang, Z. Deng, C. Finn, M. Bansal, and H. Yao, "Analyzing and mitigating object hallucination in large vision-language models," *arXiv preprint arXiv:2310.00754*, 2023.
- D.Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023.
- J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings IEEE international conference of the IEEE.Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp.
- S.Zhu, A. Kimmel, K. E. Bekris, and A. Boualiaris, "Fast model identification via physics engines for data-efficient policy search," *arXiv preprint arXiv:1710.08893*, 2017.
- X.Zhu, J. Wang, L. Zhang, Y. Zhang, R. Gan, J. Zhang, and Y. Yang, "Solving math word problem via cooperative reasoning induced language models," *arXiv preprint arXiv:2210.16257*, 2022.
- Y.Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*.IEEE, 2017, 第 3357-3364 页。

人工智能代理附录

A GPT-4V 代理提示详细信息

除非特别说明，否则我们使用 GPT-4V 的默认系统提示。我们在图 31 中详细描述了提示 GPT-4V 使用 Minecraft 的过程，以及如何生成 GPT-4V 的响应。

B 用于出血边缘的 GPT-4V

《出血边缘》是一款第三人称团队战斗游戏，玩家试图夺取目标点或收集比敌方团队更多的资源。图 32 展示了 GPT-4V 在提示游戏 "Bleeding Edge" 时的输入和输出示例。与 Minecraft 相比，我们发现 GPT-4V 对视觉内容和游戏规则的理解不够透彻。这可能是由于：（1）GPT-4V 的训练数据中存在大量 Minecraft 数据；（2）与 Minecraft 相比，Bleeding Edge 的视觉复杂性。

C 用于微软格斗模拟器的 GPT-4V

如图 33 所示，基于 GPT-4V 的代理可以为 Microsoft 飞行模拟器中的玩家提供高级动作描述。它描述了玩家如何通过驾驶舱视角和飞机外部视角驾驶飞机，管理各种飞行控制和仪器，以保持适当的空速和高度，同时在虚拟空域中航行。

D 刺客信条奥德赛》的 GPT-4V

如图 34 所示，GPT-4V 代理提供了高级动作描述，即图像中的角色在历史战役中参与夜间战斗，挥舞长矛并利用特殊能力（如 "公牛冲刺"）与多个标有等级指示器的敌人作战，同时在第三人称动作角色扮演设置中管理其健康和能力仪表。GPT-4V 还描述说，玩家正在进行一场大规模的近战，控制一名盔甲上有红色点缀的角色，并根据屏幕上的指示提示使用 "Bull Rush 能力"。环境是黑暗和炽热的，表明这是一个战场，玩家的角色正面临着多个敌人，头顶上有等级指示器和健康条。该动作是教程或战斗序列的一部分，玩家在其中学习并执行战斗能力。

E 用于《战争机器 4》的 GPT-4V

如图 35 所示，根据所提供的 10 帧视频序列，GPT-4V 代理可以对玩家在视频游戏中的行为进行大致描述：

"玩家在夜间操控一名角色在工业环境中穿行，在板条箱和墙壁等各种物体后面进行战略隐蔽。角色携带武器，与敌方实体交战。玩家使用枪械，也可能使用近战攻击，这在近距离战斗中有所体现。他们似乎会管理自己的弹药和健康状况，这可以通过屏幕上闪烁的红色指示灯等视觉提示来显示，红色指示灯通常表示

角色受到了伤害。

此外，玩家似乎在战术性地移动，在不同的掩体位置之间转换，以推进或抵御敌人。有证据表明，保护盾在某些时候会被激活，这表明玩家正在通过使用防御能力来应对来袭的炮火。角色似乎还在重新装载武器，这表明战斗中存在资源管理的元素。血液和撞击效果表明成功击中了敌人。

玩家的角色会与盟友合作，因为在附近会看到其他角色参与类似的战斗行动。环境中还有一些互动元素，如爆炸桶，玩家可以利用这些元素发挥自己的优势。

```
In [134]: n = 2.0
fps = 20
start_frame = int(950)
end_frame = int(950 + fps * n)
start = int(start_frame/fps*.0)
end = int(end_frame/14.0)
frames=base64Frames[start_frame:end_frame]
```

```
In [135]: display_handle = display(None, display_id=True)
for img in frames:
    display_handle.update(Image(data=base64.b64decode(img.encode("utf-8"))))
    time.sleep(0.025)
```



```
In [106]: with open("C:\\\\Users\\\\_\\\\Downloads\\\\Video\\\\Video\\\\AC0\\\\tutorial_level_AC01_inputs.json") as f:
    actions = json.load(f)
    # print(actions[100:102])
    min_ticks = start
    max_ticks = end
    action_frames = [d for d in actions if min_ticks <= d['ticks'] <= max_ticks]
```

```
In [141]: newButtons:[], "keyboard":{ "keys": [ "key.keyboard.d", "key.keyboard.w"], "newKeys":[], "chars": ""}, "isGuiOpen":false, "isGuiInventory": false
```

```
In [146]: MESSAGES = [
    {
        "role": "system",
        "content": [
            "You are playing minecraft, please generate future actions."
        ],
        "model": "gpt-4-vision-preview",
        "messages": PROMPT_MESSAGES,
        "tokens": 2000,
        "temperature": 0.7
    }
]
```

```
In [147]: result = client.chat.completions.create(**params)
print(result.choices[0].message.content)
```

Since you've collected wood, the next steps to craft a crafting table in Minecraft are as follows:

1. **Open your inventory** ('E' key by default) to bring up the crafting screen.
2. **Place the wood logs** you've collected into one of the crafting slots to turn them into wooden planks. Each log will create 4 wooden planks.
3. **Collect the wooden planks** that you've created by dragging them into your inventory.
4. **Fill all four crafting slots** in the 2x2 crafting grid (available in your inventory screen) with wooden planks to create one crafting table.
5. **Collect the crafting table** by dragging it into your inventory.
6. Close your inventory and **select the crafting table** in your hotbar.
7. **Right-click on the ground** to place the crafting table.

Once the crafting table is placed, you can interact with it to use the 3x3 crafting grid for creating a wide array of more complex items. That's the basic process for crafting a crafting table in Minecraft. Now you're ready to begin crafting more advanced items and tools.

图 31：我们提供代码片段来展示调用 GPT-4V 播放 Minecraft 视频的完整提示过程。首先，我们对视频帧进行编码，然后调用 GPT-4V 端点。模型的响应显示在屏幕下方。

GPT-4-Vision**提示：**

这些是 "出血边缘" 球员的视频帧，从左到右、从上到下排成一个网格。用一句话简单而准确地描述玩家在做什么。请具体描述重要的项目、实体和动作。在描述中不要提及具体的帧数或游戏名称。

视频输入**输出：**

玩家首先要地图上奔跑，经过不同的检查点，并与几个夺取点互动，然后与敌方玩家战斗，最后在被另一个敌人攻击的同时夺取一个目标。

图 32：GPT-4V 能够为《Bleeding Edge》等具有第三人称视角和视觉复杂场景的游戏生成有意义的内容。为了向 GPT-4V 输入大量帧（48 帧），我们以网格形式输入帧，并在每一帧上叠加帧号（如上图所示）。

GPT-4-Vision**Prompt:**

These are frames of a video of a Flight Simulator gaming player ordered from left to right and top to bottom. Give a simple, but precise description of what the player is doing in 1 sentence. Be specific about important items, entities, and actions. In your description do not mention specific frame numbers or the name of the game.

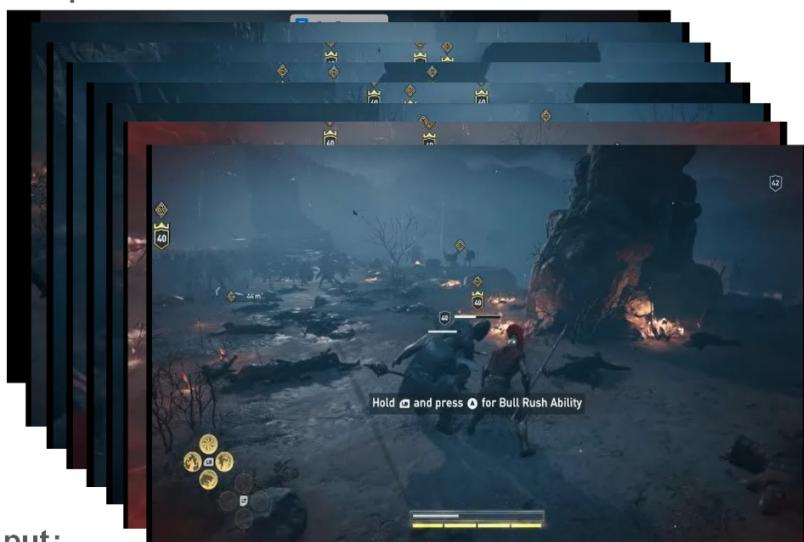
Video input:**Output:**

In the images from the flight simulator game, the player is piloting an aircraft, taking off from an airstrip over water, adjusting the cockpit controls, and navigating through the game's interface to manage the flight, while dealing with an alert about low airspeed and the need to increase it to avoid a stall.

图 33：GPT-4V 能够为微软战斗模拟器游戏提供有意义的内容描述。GPT-4V 代理可以提供高层次的动作描述，因为飞行模拟器中的玩家正在驾驶飞机，通过驾驶舱视角和飞机外部视角显示，玩家在虚拟空域中航行时，需要管理各种飞行控制和仪器，以保持适当的空速和高度。为了输入大型视频，我们选择了一些关键帧（6 个）发送给 GPT-4V。我们单独输入每个关键帧，不使用网格（如上图所示）。

GPT-4-Vision**Prompt:**

These are frames of a video of an ASSASSIN's CREED ODYSSEY gaming player ordered from left to right and top to bottom. Give a simple, but precise description of what the player is doing. Be specific about important items, entities, and actions. In your description do not mention specific frame numbers or the name of the game.

Video input:**Output:**

The player is engaged in a large-scale melee combat scenario, controlling a character with red accents on their armor, and is prompted to use a "Bull Rush Ability" as indicated by the on-screen instructions. The environment is dark and fiery, suggesting a battlefield, and the player's character is facing multiple enemies marked with level indicators and health bars overhead. The action is part of a tutorial or combat sequence where the player learns and executes combat abilities.

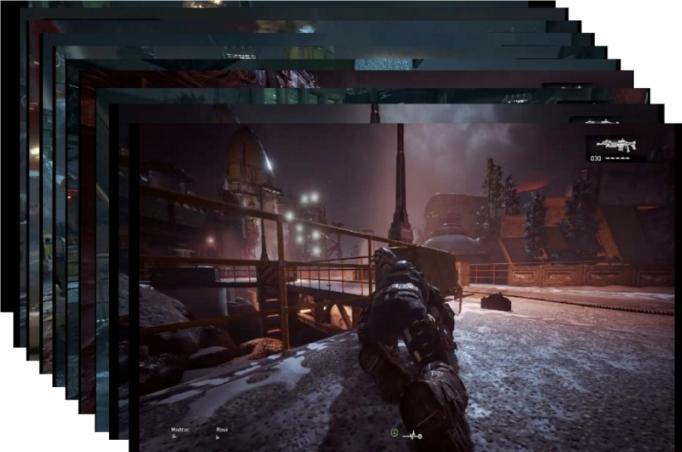
图 34：GPT-4V 能够为游戏《刺客信条奥德赛》生成有意义的内容描述。基于 GPT-4V 的代理还能提供高级动作描述，即 “在一场历史战役中，图像中的角色[正在]进行夜间战斗，在第三人称动作角色扮演设置中，挥舞长矛并利用特殊能力（如公牛冲刺）与多个标有等级指示器的敌人战斗，同时管理他们的健康值和能

力值”。为了输入大型视频，我们选择了一定数量的关键帧（8个）发送到GPT-4V。我们单独输入每个关键帧，而不使用网格（如上图所示）。

GPT-4-Vision

Prompt:
These are frames of a video of the GEARS of WAR 4 gaming player ordered from left to right and top to bottom. Give a simple, but precise description of what the player is doing. Be specific about important items, entities, and actions. In your description do not mention specific frame numbers or the name of the game.

Video input:



Output:

Based on the sequence of images provided, here is a general description of the player's actions in a video game:

The player is navigating a character through an industrial environment at night, taking strategic cover behind various objects like crates and walls. The character is armed and engages in combat with enemy entities. The player uses firearms and possibly melee attacks, as indicated by close-quarters combat. They appear to manage their ammunition and health, indicated by visual cues such as a flashing red indicator on the screen, which often signifies damage to the character.

Additionally, the player seems to be moving tactically, shifting between different cover positions to advance or defend against the enemy. There is evidence of a protective shield being activated at some point, which suggests the player is responding to incoming fire by using defensive abilities. The character also seems to be reloading their weapon, suggesting an element of resource management during combat. Blood and impact effects indicate successful hits on the enemies.

The player's character collaborates with allies, as other characters are seen in the vicinity, engaging in similar combat actions. The environment has interactive elements, such as exploding barrels, which the player might use to their advantage.

The scenes depict fast-paced action with a focus on strategic movement, shooting, and situational awareness to overcome adversaries and possibly achieve a specific objective within the game.

图 35：GPT-4V 能够为《战争机器 4》（GEARS of WAR 4）游戏生成有意义的内容描述。GPT-4V 代理提供的高级动作描述是，玩家在一个视觉丰富的未来主义环境中参与战斗，让人联想到工业或军事综合体。角色身着盔甲，配备枪支，在这个区域内穿梭，利用掩体和瞄准系统与敌人交战。捕捉到的动作包括玩家角色从掩体后瞄准、向敌人开火、根据屏幕上的视觉提示受到伤害以及进行近距离格斗。环境以夜间照明为标志，人工光源突出显示感兴趣的区域或行动区域。有时，玩家的角色会以透明的轮廓出现，这可能表示一种特殊能力或临时状态，如无敌或隐身。为了输入大型视频，我们选择了一些关键帧（10 个）发送到 GPT-4V。我们单独输入每个关键帧，而不使用网格（如上图所示）。

这些场景描绘了快节奏的动作，重点是战略移动、射击和态势感知，以战胜对手并可能实现游戏中的特定目标"。

F 星域 GPT-4V

如图 36 所示，根据为《星际争霸》提供的 10 帧视频序列，GPT-4V 代理提供了玩家在视频游戏中行动的一般描述。GPT-4V 表示："这些图像描述了《星际战场》游戏中的一系列事件，展示了玩家在采矿环境中的互动。以下是该序列的描述：

- 1) 玩家站在一个光线昏暗、钟乳石林立的地下洞穴中，身边还有一个穿着类似太空装备的角色。他们面对着一个突出显示的航点，该航点表示一个目标或感兴趣的地点。
- 2) 玩家穿过一个工业内部，可能是一个采矿站，接近另一个角色。HUD（平视显示器）显示了 "林主管" 的对话，这表明玩家与 NPC（非玩家角色）进行了互动，也可能是任务简报或更新。
- 3) 玩家似乎正跟随 NPC "林主管" 在洞穴中穿行，这暗示着 NPC 将带领玩家前往特定地点或完成特定任务。
- 4) 林主管的对话提示了一个决策点或任务更新，可能会将玩家引向一个与资源 "大矿脉" 有关的新目标。
- 5) 玩家会再次遇到林主管，他将提供有关金属矿藏的指导或信息，这意味着采矿任务或新区域探索的开始。
- 6) 在这个画面中，玩家似乎在观察一台机器或与之互动，而另一个角色伊莎贝尔则在示意与设备操作或技术工作有关的动作或对话。
- 7) 玩家俯视崎岖的地形，上面有两个发光的物体，可能代表要收集的资源，也可能代表任务的兴趣点。
- 8) 玩家要检查一种特定的资源，HUD 标记为 "镀矿床"，表明玩家正在进行资源采集或采矿活动。
- 9) 玩家在洞穴中穿行，可能是寻找更多资源，也可能是根据前方另一个身穿太空装备的角色的指示，前往下一个目标。
- 10) 玩家被安置在洞穴中一个较暗的区域，两个发光标记指示着一条路径或需要到达的目标点。

在这些图像中，玩家似乎正在执行一项任务，包括探索、与 NPC 互动、收集资源以及可能在采矿环境中操作机械。HUD 元素，如航点、对话文本和任务更新，为玩家的行动和游戏目标提供了背景信息"。

GPT-4-Vision

Prompt:
These are frames of a video of the Starfield Console game player ordered from left to right and top to bottom. Give a simple, but precise description of what the player is doing. Be specific about important items, entities, and actions. In your description do not mention specific frame numbers or the name of the game.

Video input:



Output:
The images depict a series of events within the game Starfield, showcasing the player's interactions within a mining environment. Here's a description of the sequence:
1.The player stands in a dimly lit underground cavern with stalactites, accompanied by another character in similar space gear. They face a highlighted waypoint, indicating an objective or location of interest.
2.Moving through an industrial interior, likely a mining station, the player approaches another character. The HUD (Heads-Up Display) shows a dialogue line from "Supervisor Lin," which suggests NPC (non-player character) interaction and possibly a mission briefing or update.
3.The player seems to be following the NPC, Supervisor Lin, through the cavern, hinting at a guided segment of gameplay where the NPC leads the player to a specific location or task.
4.A dialogue from Supervisor Lin suggests a decision point or mission update, possibly directing the player to a new objective related to a "big vein" of resources.
5.The player encounters Supervisor Lin again, who provides instruction or information about metal deposits, implying the start of a mining task or exploration of a new area.
6.In this frame, the player appears to be observing or interacting with a piece of machinery, with another character, Isabelle, indicating an action or dialogue related to equipment operation or technical work.
7.The player looks down at a rough terrain with two glowing objects, which could represent resources to be collected or points of interest for the mission.
8.The player examines a specific resource, indicated by a HUD marker as "Beryllium Deposit," suggesting a resource gathering or mining activity.
9.The player traverses through the caverns, potentially searching for additional resources or navigating to the next objective, as indicated by another character in space gear ahead.
10.The player is positioned in a darker area of the cavern, with two glowing markers indicating a path or objective points to be reached.
Throughout these images, the player seems to be engaged in a mission involving exploration, NPC interaction, resource collection, and possibly machinery operation within a mining environment. The HUD elements like waypoints, dialogue text, and mission updates provide context to the player's actions and the game's objectives.

图 36：GPT-4V 能够为 X-Box 游戏机游戏 "星域" 提供有意义的内容描述。GPT-4V 代理为这款太空主题探索游戏提供了高级动作描述。为了输入大型视频，我们选择了一些关键帧（10 帧）发送给 GPT-4V。我们单独输入每个关键帧，不使用网格（如上图所示）。

作者简介

Zane Durante 斯坦福大学计算机科学博士生 电子邮箱:

durante@stanford.edu

网址: <https://zanedurante.github.io/>

Zane Durante 是斯坦福视觉实验室的博士生，由李飞飞博士指导。他的研究兴趣包括视频理解、计算机视觉与自然语言的连接以及人工智能在医院护理中的应用。他正在领导斯坦福医学院和斯坦福视觉实验室之间的合作，开发一个医院场景视频剪辑自然语言描述的数据集。他的博士研究工作得到了国家自然科学基金研究生研究奖学金的支持。

Qiuyuan Huang 美国华盛顿州雷德蒙德微软研究院首席研究员。电子邮件:

qihua@microsoft.com

网址: <https://www.microsoft.com/en-us/research/people/qihua/>

黄秋媛是华盛顿州雷德蒙德微软研究院（MSR）深度学习小组的首席研究员。她目前的研究兴趣主要集中在深度学习、多模态和自然语言处理领域，特别是用于游戏、机器人和医疗保健的代理人工智能；用于交互式人工智能的知识推理智能；用于推理的神经符号计算；以及用于 NLP 和多模态的大型基础模型。

Naoki Wake 美国华盛顿州雷德蒙德微软公司研究员。电子

邮件: naoki.wake@microsoft.com

网址: <https://www.microsoft.com/en-us/research/people/nawake/>

Naoki Wake 是微软公司应用机器人研究小组的研究员。他目前的研究涉及机器人多模态感知系统和协同语音手势系统的开发。他过去的研究涉及听觉神经科学、神经康复和语音处理。Naoki 于 2014 年获得东京大学工程学学士学位，并于 2019 年获得信息科学与技术博士学位。

Ran Gong 加州大学洛杉矶分校计算机科学博士生。电子邮件: nikepupu@ucla.edu

网址: <https://nikepupu.github.io>

Ran Gong 是加州大学洛杉矶分校 VCLA 实验室的博士生。他的研究领域涉及机器人学、计算机视觉、计算机图形学和机器学习。他的研究重点是具身模拟和交互，目标是创造能在不同环境中解决不同任务的智能行为，以及与人类协作的能力。他在加州大学洛杉矶分校获得计算机科学与工程学士学位。

Jae Sung Park 华盛顿大学博士生 电子邮箱:

jspark96@cs.washington.edu

网站: <https://homes.cs.washington.edu/~jspark96/>

Jae Sung 是一名博士生，导师是 Yejin Choi 和 Ali Farhadi。他的研究重点是开发具有多模态常识推理能力的模型。他感兴趣的是为模型配备将语言概念与视觉模式相结合的基础，并让模型以人类处理视觉信息的方式理解多媒体内容。Jae Sung 在加州大学伯克利分校获得计算机科学学士学位。

Bidipta Sarkar 斯坦福大学本科生 电子邮箱:

bidiptas@stanford.edu

网址: <https://bsarkar321.github.io/>

比迪普塔-萨卡尔 (Bidipta Sarkar) 是斯坦福大学的一名大四本科生，也是斯坦福 ILIAD 实验室的成员。他的研究重点是创建能与环境互动并能安全地与人类和其他自主代理一起工作的人工智能代理。

Rohan Taori 斯坦福大学计算机科学博士生 电子邮箱：

rtaori@cs.stanford.edu

网址： <https://www.rohantaori.com/>

Rohan Taori 是斯坦福人工智能实验室的一名博士生。他的研究重点是研究现实世界系统中的机器学习基础。最近，Rohan 推动了开源大型语言模型的发展，对其进行微调，使其成为有用的通用助手。此外，罗汉还对增强语言模型的多模态能力非常感兴趣，使其能够对图像和视频进行推理。Rohan 在加州大学伯克利分校获得计算机科学学士学位。

Yusuke Noda 首席软件工程师。华盛顿州雷德蒙德微软游戏公司。电子邮件：

yusuke.noda@microsoft.com

网址：<https://www.linkedin.com/in/yusuke-noda-908797/>

野田雄介是微软游戏平台的首席软件工程师。他领导了 Xbox One 和 Xbox Cloud Gaming 游戏基础架构的开发，拥有超过 16 年的云计算和游戏技术高效基础架构开发经验。

Demetri Terzopoulos 加州大学洛杉矶分校特聘教授、奥斯卡获奖计算机科学家和企业家。电子邮件：

dt@cs.ucla.edu

网站：<https://web.cs.ucla.edu/~dt/>

Demetri Terzopoulos 博士是加州大学洛杉矶分校计算机科学的特聘教授和校长教授，并担任加州大学洛杉矶分校计算机图形与视觉实验室主任。他还是跨国医疗人工智能公司 VoxelCloud, Inc. 的联合创始人兼首席科学家。他现在或曾经是古根海姆研究员、伦敦皇家学会（FRS）和加拿大皇家学会（FRSC）研究员、美国计算机协会（ACM）研究员、电气和电子工程师学会（IEEE）终身研究员、亚太人工智能协会（AAIA）研究员、国际工程技术学会（IETI）杰出研究员、欧洲科学院（EAS）和纽约科学院（NYAS）院士，以及 Sigma Xi 终身会员。

Yejin Choi 华盛顿大学 Wisnner-Slivaka 讲座教授兼 Brett Helsel 教授、艾伦人工智能研究所高级研究经理和麦克阿瑟研究员。

电子邮件：yejin@cs.washington.edu

网站：<https://homes.cs.washington.edu/~yejin/>

叶瑾是华盛顿大学维斯纳-斯利瓦卡讲座教授和布雷特-赫尔塞尔教授，以及艾伦人工智能研究所高级研究经理。她曾于 2018 年获得安妮塔-博格早期职业奖。她曾于 2020 年获得麦克阿瑟基金会奖学金。她曾获得 AAAI 2020、Neurips 2021、ICML 2022 和 ACL 2023 的优秀论文奖，以及 NAACL 2022 和 ACL 2023 的最佳论文奖。她是 COLM 2024 的主要组织者之一，COLM 2024 是一个专注于语言建模研究的学术场所。她的主要研究兴趣是自然语言处理、机器学习和人工智能领域的常识推理，并对计算机视觉和数字人文有更广泛的兴趣。

李飞飞 斯坦福大学计算机科学教授 电子邮件：

feifeili@stanford.edu

网址：<https://profiles.stanford.edu/fei-fei-li/>

菲菲是斯坦福大学计算机科学系首任红杉教授，也是斯坦福以人为核心的人工智能研究所的联合主任。2013 年至 2018 年，她担任斯坦福大学人工智能实验室主任。而在 2017 年 1 月至 2018 年 9 月从斯坦福大学休假期间，李博士担任谷歌副总裁，并担任谷歌云 AI/ML 首席科学家。此后，她在多家上市或私营公司担任董事会成

员或顾问。

Katsushi Ikeuchi 微软公司高级首席研究经理，IEEE 终身会士。微软研究院，华盛顿州雷德蒙德。IEEE 终身会士。

电子邮件：katsuike@microsoft.com

网址：<https://www.microsoft.com/en-us/research/people/katsuike/>

Katsushi Ikeuchi 博士 1973 年获得京都大学机械工程学士学位，1978 年获得东京大学信息工程博士学位。曾在麻省理工学院人工智能实验室工作三年，在日本政府通商产业省电工实验室工作五年，在卡内基梅隆大学机器人研究所工作十年、

在东京大学工业科学研究所工作 19 年后，他于 2015 年加入微软担任首席研究员。在 CMU 和东京大学任职期间，他指导了 50 多名博士生。

他的研究兴趣横跨计算机视觉、机器人和计算机制图。在这些研究领域，他曾多次获得最佳论文奖，包括计算视觉领域的戴维-马尔奖（David Marr Prize）和电气和电子工程师协会机器人与自动化学会的傅国新纪念最佳交易论文奖。

他的社会服务包括：IROS95、ITSC99、IV01、ICCV05、ACCV07、ICCV17 的总主席；CVPR96、ICCV03、ICRA09、ICPR12、ICCV15 的项目主席；IJCV 的 EIC（2000-2017），IJ ITS 的 EIC（2012-2014），副主席

IEEE Trans.RA、IEEE Trans.PAMI 编辑；2000-2002 年任 IEEE 信号处理学会杰出讲师，2004-2006 年任 IEEE 机器人与自动化学会杰出讲师，2008-2010 年任 IEEE 计算机学会杰出讲师。

通过这些研究和社会服务，他被 IEEE、IEICE、IPSJ 和 RSJ 授予终身研究员。他还获得了 IEEE-PAMI 颁发的杰出研究员奖、日本天皇颁发的紫绶荣誉勋章、大川财团颁发的大川奖以及日本专业协会颁发的多项研究成就奖。

Hoi Vo 技术研究员。微软游戏和 X-Box 新兴技术部，华盛顿州雷德蒙德。电子邮件：

hoiv@microsoft.com

网址：<https://www.linkedin.com/in/hoi-vo-193420/>

Hoi Vo 是微软游戏部门的技术研究员，专注于连接 AGI 与新游戏 IP。他在建立 Xbox 云游戏服务方面发挥了关键作用，并领导了微软各团队的云和游戏工作，包括微软的 Windows Azure 平台。他的工作重点是优化人工智能模型，使其在边缘高效运行，同时利用云技术实现超越硬件能力的应用场景，包括微调模型，使其更符合各种游戏体验。

高剑锋 微软公司杰出科学家、副总裁，IEEE 会员。微软研究院，华盛顿州雷德蒙德。IEEE Fellow。

电子邮件：jfgao@microsoft.com

网址：<https://www.microsoft.com/en-us/research/people/jfgao/>

高剑锋博士是微软研究院杰出科学家、副院长，IEEE Fellow，ACM 杰出会员，现任微软研究院深度学习组组长。2014年至2017年，他在微软人工智能研究院和雷德蒙德微软研究院深度学习技术中心（DLTC）担任商业人工智能合作伙伴研究经理。他领导开发了预测性销售和营销的人工智能解决方案。他还致力于文本和图像处理的深度学习（见 ACL/SIGIR 2018 Tutorial、Deep Learning 2017 Tutorial 和 IJCAI 2016 Tutorial 或 MS 内部网站），并领导开发对话、机器阅读理解（MRC）和问题解答（QA）的人工智能系统。从 2022 年开始，他领导自我完善人工智能的研究，其中 LLM（例如 ChatGPT/GPT4）被增强和调整用于开发商业人工智能系统。

致谢

我们特别感谢 Peter Lee、Doug Burger、Desney Tan、Johannes Gehrke、Ryan White、Ece Kamar、Subhojit Som 和 Kareem Choudhry 的建议、大力支持和鼓励。我们感谢张海燕、Spencer Perreault、Dave Bignell、Katja Hofmann、Sam Devlin、Shanzheng Tan、Raluca Georgescu、Bill Dolan、Nebojsa Jojic、Sudha Rao、Adrian Brown、Andrzej Banburski-Fahey、杨建伟等人早期的深入讨论和帮助。我们感谢 Kiran Muthabatulla、Antonio Criminisi、Tom Cashman、Nguyen Bach、Jennifer Marsman、Mesh 团队的 Jaron Lanier、微软的 OCTO 和微软 office 团队的混合现实工作、数据集工作以及他们对本项目的慷慨帮助和反馈。我们特别感谢 Paul Bennett、Corby Rosset、Michel Galley、王成龙、彭宝林、陈浩、Silviu Cucerzan、Ahmed Awadallah 和 Saleema Amershi 对 NLP 部分提出的建议和意见。作者衷心感谢 Paul Smolensky、Yonatan Bisk、陈可真、王博睿、桂良科、王欣、甘喆、马晓健、郑子龙、朱松春、Dragomir R. Radev、Daniel McDuff、Harry Shum 对本文的相关前人工作、评论、建议、不厌其烦的多次审阅以及他们对文献的指点。最后，我们衷心感谢微软 Hololens 团队、微软 X-box 团队和 Meta Quest 团队慷慨提供设备；MSR 中央工程（CE）团队、微软 343 团队提供数据收集和共享；微软 AOAI 和 GCR 团队提供 Azure-OpenAI 终端支持。

我们要感谢斯坦福大学 "人工智能辅助医疗伙伴关系"（Partnership in AI-assisted Care）的同事们，他们为本研究中探讨的医疗应用提供了帮助。我们尤其要感谢 Ehsan Adeli、Paul Tang、Amit Kaushal、Roger Bohn、Kevin Schulman 和 Arnold Milstein 的临床专业知识和指导。

本研究得到了微软研究院项目 Fair 2023、微软 HackBox 2023 和 OCTO 团队的支持。